이 학 박 사 학 위 논 문

Contextual multi-armed bandit algorithm for semiparametric reward model

준모수적 가법 모형을 위한 새로운 다중 슬롯 머신 알고리즘

2019년 2월

서울대학교 대학원

통계학과

김 지 수

Contextual multi-armed bandit algorithm for semiparametric reward model

준모수적 가법 모형을 위한 새로운 다중 슬롯 머신 알고리즘

지도교수 Myunghee Cho Paik

이 논문을 이학박사 학위논문으로 제출함

2018년 10월

서울대학교 대학원

통계학과

김 지 수

김지수의 이학박사 학위논문을 인준함

2018년 12월

위 원 장	이 영 조	(인)
부위원장	Myunghee Cho Paik	(인)
위 원	이 재 용	(인)
위 원	오 희 석	(인)
위 원	김 재 광	(인)

Contextual multi-armed bandit algorithm for semiparametric reward model

By

Gi-Soo Kim

A Thesis
Submitted in fulfillment of the requirement
for the degree of
Doctor of Philosophy
in Statistics

Department of Statistics
College of Natural Sciences
Seoul National University
February, 2019

ABSTRACT

Contextual multi-armed bandit algorithm for semiparametric reward model

Gi-Soo Kim
The Department of Statistics
The Graduate School
Seoul National University

Contextual multi-armed bandit (MAB) algorithms have been shown promising for maximizing cumulative rewards in sequential decision tasks under uncertainty when contextual information is given. Applications include news article recommendation systems, web page ad placement algorithms, revenue management, and mobile health. However, most of the proposed contextual MAB algorithms rely on strong, linear assumptions between the reward and the context of the action. This thesis proposes a new contextual MAB algorithm for a relaxed, semiparametric reward model that supports nonstationarity. The proposed method is less restrictive, easier to implement and faster than two alternative algorithms that consider the same model. It can be shown that the high-probability

upper bound of the regret incurred by the proposed algorithm has the same order as the Thompson sampling algorithm for linear reward models without restricting action choice probabilities. The proposed algorithm and existing algorithms are evaluated via simulation and also applied to Yahoo! news article recommendation log data provided by Yahoo! Webscope.

Keywords: Contextual multi-armed bandit algorithm, sequential decision, Thompson sampling, semiparametric model.

Student Number: 2015 − 30089

Contents

Abstract **i**

1 Introduction **1**

2 Literature Review **6**

2.1 The multi-armed bandit problem 6

2.2 Linear contextual MAB 7

2.2.1 Upper confidence bound (UCB) algorithm . 9

2.2.2 Thompson sampling (TS) algorithm 15

2.3 Adversarial MAB 20

2.3.1 EXP4.P algorithm 21

2.4 Semiparametric contextual MAB 23

2.4.1 Action-centered TS algorithm 26

2.4.2 BOSE algorithm 28

3 Proposed method **31**

3.1 Proposed algorithm 31

3.2 Proof 33

3.2.1 Stage (a) 34

3.2.2 Stage (b) 40

3.2.3 Stage (c) 42
3.2.4 Stage (d) 42
3.2.5 Stage (e) 44
3.2.6 Stage (f) 44

4 Simulation study **46**

5 Real data analysis **50**
5.1 Off-policy evaluation method 52
5.1.1 Assumptions 52
5.1.2 Algorithm : when L selects each arm uniformly at random. 52
5.1.3 Algorithm 2 : when L does not select each arm uniformly at random. 55
5.2 Application results 57

6 Concluding remarks **59**

Abstract (in Korean) **64**

List of Tables

5.1 Mean, 1st quartile (1st Q.) and 3rd quartile (3rd Q.) of user clicks achieved by each algorithm over 10 runs 58

List of Figures

4.1 Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations in case (i). . 47

4.2 Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations in case (ii). . 48

4.3 Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations in case (iii). 49

4.4 Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations in case (iv). . 49

5.1 Yahoo! Featured tab screenshot image 51

Chapter 1

Introduction

The multi-armed bandit (MAB) problem (Robbins, 1952) formulates the sequential decision problem in which a learner must choose an action among a number of selectable actions given by the environment at each step so as to maximize the cumulated rewards. The actions are often described as the arms of a bandit slot machine with multiple arms. The act of choosing an action is characterized as pulling an arm of the bandit machine, where different arms give possibly different rewards. By repeating the process of pulling arms and receiving rewards, the learner accumulates information about the reward compensation mechanism and learns from it, choosing the arm that is close to optimal as time passes. The MAB problem is a type of reinforcement learning problem. The difference with the full reinforcement learning problem is that the learner has no control on the environment. Application areas include the mobile healthcare system (Tewari and Murphy, 2017), web page ad placement algorithms (Langford et al., 2008), news article placement algorithms (Li et al., 2010),

revenue management (Ferreira et al., 2018), marketing (Schwartz et al., 2017), and recommendation systems (Kawale et al., 2015).

For example, the Yahoo! web system uses a news article recommendation algorithm to select one article among a large pool of available articles and displays it on the Featured tab of the web page every time a user visits. The user clicks the article if he or she is interested in the contents. The goal of the algorithm is to maximize the cumulated number of user clicks. After each visit, the algorithm reinforces its article selection strategy based on past user click feedback. In this setting, available articles corresponds to different actions and the user click corresponds to a reward. The challenging part of the MAB problem is that the reward of the action that the learner has not previously chosen is forever unknown, i.e., whether the user would have clicked or not the unchosen article remains missing. Therefore, the learner should balance well between "exploitation" and "exploration" at each selection step. "Exploitation" means selecting the best action that information accumulated so far points to, while "exploration" refers to choosing an action that will assist in future choices, although it does not seem to be the best option at the moment.

The MAB problem was first theoretically analyzed by Lai and Robbins (1985). The algorithms widely used in mobile healthcare systems or ad and news article placement algorithms are of a more extended form, called contextual MAB algorithms. A contextual MAB algorithm enables at each selection step the use of side information about each action, called context, given in the form of finite-dimensional covariates. For example, in the news article rec-

ommendation example, information on the visiting user as well as the articles are given in the form of context vectors. In 2010, the Yahoo! team (Li et al., 2010) proposed a contextual MAB algorithm that achieved a 12.5% click lift compared to a context-free MAB algorithm. Still, the method of Li et al. (2010) and other existing algorithms rely on rather strong assumptions on the distribution of the reward. In particular, most of the existing algorithms assume that the expectation of the reward of a particular action has a time-invariant, linear relationship with the context vector. This assumption can be restrictive in real world settings where the rewards typically adapt to past actions.

In this thesis, we propose a novel contextual MAB algorithm which works well under a relaxed assumption on the distribution of rewards. This relaxed assumption supports nonstationarity of the reward via an additive intercept term to the original time-invariant linear term. This intercept term changes with time but does not depend on the action. We enable consistent estimation of the regression parameter in the linear term using a centering method on context vectors. We prove using novel martingale inequalities that under the semiparametric reward model, the high probability upper bound of the difference between the maximum possible mean of expected rewards and the mean of the expected rewards incurred by the proposed algorithm decreases to 0 at the same rate achieved by the existing algorithms but under more restrictive linear assumptions.

Alternative methods have been proposed for the same reward model by Greenewald et al. (2017) and Krishnamurthy et al. (2018).

The performance of the first method is guaranteed under restrictive conditions on the action choice probabilities, whereas the second method is computationally heavy since it requires $O(N^2)$ computations at each iteration where N denotes the number of arms. Moreover, Krishnamurthy et al. (2018) did not specify the action selection distribution when $N > 2$. Our method improves on the previous results in that it does not restrict action choice probabilities, requires $O(N)$ computations, and specifies the action selection distribution for every N. Furthermore, the proposed estimator for the regression parameter achieves fastest convergence rate to the true parameter.

The rest of the thesis is organized as follows. In Chapter 2, we review existing contextual bandit algorithms and their theoretical properties. In Chapter 3, we present a new contextual MAB algorithm which works well under a semiparametric reward model which supports nonstationarity. We also present a new theorem on the theoretical performance of the proposed method. Then we evaluate the proposed method and existing methods via simulation in Chapter 4. In Chapter 5, we apply the proposed method to Yahoo! news article recommendation log data provided by Yahoo! Webscope. This data was obtained by applying an uniform random policy for choosing which article to place on the Featured tab. Evaluating a new reinforcement learning policy retrospectively using observational real data is a challenging task itself because in the data, the rewards of the actions that were not chosen by the original policy are missing. This problem is called off-policy evaluation problem. In Chapter 5, we applied the off-policy evaluation

method of Li et al. (2011) to unbiasedly estimate the expected total reward incurred by the proposed algorithm. We first discuss the method of Li et al. (2011) and then present the application results. Finally, concluding remarks follow in Chapter 6.

Chapter 2

Literature Review

2.1 The multi-armed bandit problem

The classic MAB problem considers the case where the learner is repeatedly faced with N possible actions which yield rewards from possibly different distributions. Specifically, we assume that at time t, the i-th arm ($i = 1, \cdots, N$) yields a random reward $r_i(t)$ with unknown mean θ_i, i.e., $\mathbb{E}(r_i(t)) = \theta_i$. Among the N arms, the learner pulls one arm $a(t)$, and observes reward $r_{a(t)}(t)$. Under such settings, let $\theta^* = \max_{1 \le i \le N} \theta_i$ and $regret(t)$ be the difference between the expected reward of the optimal arm and the expected reward of the arm chosen by the learner at time t, i.e.,

$$regret(t) = \theta^* - \mathbb{E}(r_{a(t)}(t)) = \theta^* - \theta_{a(t)}.$$

Then, the goal of the learner is to minimize the sum of regrets over T steps defined as follows,

$$R(T) := \sum_{t=1}^{T} regret(t) = \theta^* T - \sum_{t=1}^{T} \theta_{a(t)}.$$

To minimize $R(T)$, the learner has to learn the values of θ_i's. Since only $r_{a(t)}(t)$ is observed among the whole reward vector $r(t) = [r_1(t), \cdots, r_N(t)]^T$, the learner should balance between exploitation and exploration.

Lai and Robbins (1985) proposed an adaptive allocation rule which selects actions based on a sharp upper confidence bound of each θ_i. The rule pulls the arm which has the highest upper confidence bound. The upper confidence bound reflects both the current estimate of θ_i and its uncertainty. Therefore, the rule pulls an arm if either the reward estimate is high (exploitation) or the uncertainty of the estimate is high (exploration). The upper confidence bound rule can also be viewed as an application of the principle of optimism in the face of uncertainty. The upper confidence bound is an optimistic guess of each action, and the learner pulls the arm with the highest guess. In the same paper, Lai and Robbins (1985) proved that under some mild conditions on the reward distributions, the regret of their algorithm asymptotically achieves the lower bound $O(\log(T))$.

2.2 Linear contextual MAB

In the aforementioned classic MAB problem, each arm has a fixed reward distribution, not changing over time. However in some real settings such as the news article recommendation example, different users have different tastes and tendencies to click a certain article. Hence, the reward mechanism of a certain arm differs according to the incoming user and article characteristics. In the contextual MAB problem, we assume that there is a finite-dimensional

context vector $b_i(t) \in \mathbb{R}^d$ associated with each arm i at time t and that $r_i(t)$ depends on $b_i(t)$, i.e.,

$$\mathbb{E}(r_i(t)|b_i(t)) = \theta(b_i(t)), \quad i = 1, \cdots, N,$$

where $\theta(\cdot)$ is an arbitrary function. Specifically, linear contextual MAB problems assume that $\theta(b_i(t))$ is linear in $b_i(t)$,

$$\mathbb{E}(r_i(t)|b_i(t)) = b_i(t)^T\mu, \quad i = 1, \cdots, N, \tag{2.1}$$

where $\mu \in \mathbb{R}^d$ is unknown. Under (2.1), the optimal arm changes over time. Let the optimal arm at time t be $a^*(t) := \underset{1\le i\le N}{\operatorname{argmax}}\{b_i(t)^T\mu\}$. Then, the regret at time t is defined as

$$\begin{aligned} regret(t) &= \mathbb{E}\big(r_{a^*(t)}(t) - r_{a(t)}(t) \;\big|\; \{b_i(t)\}_{i=1}^N, a(t)\;\big) \\ &= b_{a^*(t)}(t)^T\mu - b_{a(t)}(t)^T\mu. \end{aligned}$$

We additionally assume that given $b_i(t)$, $\eta_i(t) := r_i(t) - \mathbb{E}(r_i(t)|b_i(t))$ is R-sub-Gaussian for some $R > 0$, i.e., for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda\eta_i(t))|b_i(t)] \le \exp(\frac{\lambda^2R^2}{2}). \tag{2.2}$$

Note that this assumption is satisfied whenever $r_i(t) \in [b_i(t)^T\mu - R, b_i(t)^T\mu + R]$. Without loss of generality, we assume $||b_i(t)||_2 \le 1$ and $||\mu||_2 \le 1$, where $||\cdot||_p$ denotes the L_p-norm.

Dani et al. (2008) proved that for any algorithm, when the number of arms N is allowed to be infinite, there exists a distribution of contexts and rewards such that $R(T)$ is of order $\Omega(d\sqrt{T})$. When N is finite and $d^2 \le T$, Chu et al. (2011) showed a lower bound of $\Omega(\sqrt{dT})$. We note that the lower bounds do not depend on N but only on the dimension d of μ. Also, no algorithm can achieve better rate than $O(\sqrt{T})$ in terms of T.

2.2.1 Upper confidence bound (UCB) algorithm

Auer (2002), Li et al. (2010) and Chu et al. (2011) proposed an upper confidence bound (UCB) algorithm for the linear contextual MAB problem. We present here the algorithm of Li et al. (2010) and Chu et al. (2011).

Algorithm 1 UCB algorithm (Li et al., 2010; Chu et al., 2011)

Set $\alpha > 0$, $B = I_d$, $y = 0_d$.
for $t = 1, \cdots, T$ **do**
  Compute $\hat{\mu}(t) = B^{-1}y$.
  for $i = 1, \cdots, N$ **do**
    Compute $U_i(t) = b_i(t)^T \hat{\mu}(t) + \alpha s_{t,i}$,
    where $s_{t,i} = \sqrt{b_i(t)^T B^{-1} b_i(t)}$.
  end for
  Pull arm $a(t) = \underset{1 \le i \le N}{\operatorname{argmax}}\, U_i(t)$ and get reward $r_{a(t)}(t)$.
  Update B and y:
  $B \leftarrow B + b_{a(t)}(t) b_{a(t)}(t)^T$, $y \leftarrow y + b_{a(t)}(t) r_{a(t)}(t)$.
end for

In Algorithm 1, suppose that $U_i(t)$ is a high-probability upper bound of $b_i(t)^T \mu$, i.e., for some $\alpha > 0$ and $\forall \delta \in (0, 1)$, the following holds with probability at least $1 - \delta$:

$$|b_i(t)^T (\hat{\mu}(t) - \mu)| \le \alpha s_{t,i}, \tag{2.3}$$

for all $i = 1, \cdots, N$ and all $t = 1, \cdots, T$. Under (2.3),

$$\begin{aligned} b_{a^*(t)}(t)^T \mu &\le b_{a^*(t)}(t)^T \hat{\mu}(t) + \alpha s_{t,a^*(t)} \\ &\le b_{a(t)}(t)^T \hat{\mu}(t) + \alpha s_{t,a(t)} \\ &\le b_{a(t)}(t)^T \mu + 2\alpha s_{t,a(t)}, \end{aligned}$$

where the first and third inequalities follow from (2.3) and the second inequality is due to the action selection mechanism of Algorithm 1. Hence when (2.3) is true, we can bound $regret(t)$ for all t with high probability. With probability at least $1-\delta$,

$$regret(t) = b_{a^*(t)}(t)^T\mu - b_{a(t)}(t)^T\mu \leq 2\alpha s_{t,a(t)}, \tag{2.4}$$

for all $t = 1, \cdots, T$. Let $B(t)$ be the matrix B at time t, i.e., $B(t) = I_d + \sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T$. To bound $\sum_{t=1}^{T} s_{t,a(t)}$, Abbasi-Yadkori et al. (2011) established the following lemma.

Lemma 2.2.1. *(Lemma 11 of Abbasi-Yadkori et al., 2011) Let $\{X_t\}_{t=1}^T$ be a sequence in $\mathbb{R}^d$ with $||X_t||_2 \leq 1$, Q a $d \times d$ positive definite matrix with $det(Q) \geq 1$ and $A(t) = \sum_{\tau=1}^{t-1} X_\tau X_\tau^T$. Then, we have*

$$\sum_{t=1}^{T} X_t^T \{Q + A(t)\}^{-1} X_t \leq 2\log\Big(\frac{det(Q + A(T+1))}{det(Q)}\Big).$$

From this lemma, we can derive,

$$\sum_{t=1}^{T} s_{t,a(t)} \leq \sqrt{2dT\log(1 + T/d)}. \tag{2.5}$$

Proof. Take $X_t = b_{a(t)}(t)$, $Q = I_d$, and $A(t) = \sum_{\tau=1}^{t-1} X_\tau X_\tau^T$. Then

by Jensen's inequality and Lemma 2.2.1,

$$\begin{aligned}
\sum_{t=1}^{T} s_{t,a(t)} &\leq \sqrt{T\sum_{t=1}^{T} s_{t,a(t)}^2} \quad (\because \text{ Jensen's inequality}) \\
&= \sqrt{T\sum_{t=1}^{T} X_t^T B(t)^{-1} X_t} \\
&= \sqrt{T\sum_{t=1}^{T} X_t^T \{Q + A(t)\}^{-1} X_t} \\
&\leq \sqrt{2T\log\Big(\frac{det(Q + A(T+1))}{det(Q)}\Big)} \quad (\because \text{ Lemma 2.2.1}) \\
&\leq \sqrt{2dT\log\Big(1 + \frac{T}{d}\Big)}.
\end{aligned}$$

The last inequality is due to the determinant-trace inequality,

$$det(B(T+1)) \leq \Big(\frac{trace(B(T+1))}{d}\Big)^d \leq \Big(1 + \frac{T}{d}\Big)^d.$$

□

Summing both sides of (2.4) over t and using (2.5), we have with probability at least $1 - \delta$,

$$R(T) \leq 2\alpha \sum_{t=1}^{T} s_{t,a(t)} \leq 2\alpha\sqrt{2dT\log(1 + T/d)}. \tag{2.6}$$

Note that if $\alpha = o(\sqrt{T/\log(T)})$, the high-probability upper bound of $R(T)$ is sublinear in T, which is a desired property for an online learning algorithm.

Under the assumption that the observed rewards $r_{a(1)}(1), \cdots, r_{a(T)}(T)$ are independent given $b_{a(1)}(1), \cdots, b_{a(T)}(T)$, Chu et al. (2011) proved that (2.3) holds when $\alpha = 2R\sqrt{\log(\frac{2NT}{\delta})} + 1$. Let

w be a vector in $\mathbb{R}^{(t-1)}$ such that $w_\tau = b_i(t)^T B(t)^{-1} b_{a(\tau)}(\tau)$, for $\tau = 1, \cdots, t-1$. Then we have, for $\forall \varepsilon > 0$ and $\forall \lambda \in \mathbb{R}$,

$$\begin{aligned}\mathbb{P}(b_i(t)^T(\hat{\mu}(t) - \mu) > \varepsilon + s_{t,i}) &\leq \mathbb{P}\Big(\sum_{\tau=1}^{t-1} w_\tau \eta_{a(\tau)}(\tau) > \varepsilon\Big) \\ &\leq \mathrm{e}^{(-\lambda\varepsilon/R)}\mathbb{E}\Big(\exp\Big(\lambda \sum_{\tau=1}^{t-1} w_\tau \frac{\eta_{a(\tau)}(\tau)}{R}\Big)\Big),\end{aligned}$$

where the first inequality is due to the fact that $b_i(t)^T(\hat{\mu}(t) - \mu) = \sum_{\tau=1}^{t-1} w_\tau \eta_{a(\tau)}(\tau) - b_i(t)^T B(t)^{-1}\mu \leq \sum_{\tau=1}^{t-1} w_\tau \eta_{a(\tau)}(\tau) + s_{t,i}$, and the second inequality is the Chernoff inequality. If we assume that w is fixed and $\eta_{a(\tau)}(\tau)$'s are independent, we can proceed:

$$\begin{aligned}\mathbb{P}(b_i(t)^T(\hat{\mu}(t) - \mu) > \varepsilon + s_{t,i}) &\leq \mathrm{e}^{(-\lambda\varepsilon/R)} \prod_{\tau=1}^{t-1} \mathbb{E}\Big(\exp\Big(\lambda w_\tau \frac{\eta_{a(\tau)}(\tau)}{R}\Big)\Big) \\ &\leq \mathrm{e}^{(-\lambda\varepsilon/R)} \prod_{\tau=1}^{t-1} \exp\Big(\frac{\lambda^2 w_\tau^2}{2}\Big) \\ &= \mathrm{e}^{(-\lambda\varepsilon/R)} \exp\Big(\frac{\lambda^2 ||w||_2^2}{2}\Big) \\ &= \exp\Big(-\frac{\varepsilon^2}{2R^2||w||_2^2}\Big) \\ &\quad \Big(\because \text{take } \lambda = \varepsilon/(R||w||_2^2)\Big) \\ &\leq \exp\Big(-\frac{\varepsilon^2}{2R^2 b_i(t)^T B(t)^{-1} b_i(t)}\Big) \\ &= \exp\Big(-\frac{\varepsilon^2}{2R^2 s_{t,i}^2}\Big), \qquad (2.7)\end{aligned}$$

where the second inequality follows from the R-sub-Gaussian assumption. Taking the last term to be equal to $\frac{\delta}{2Nt^2}$ for $\forall \delta \in (0,1)$, we have with probability at least $1 - \frac{\delta}{2}$,

$$b_i(t)^T(\hat{\mu}(t) - \mu) \leq 2Rs_{t,i}\sqrt{\log(\frac{2Nt}{\delta})} + s_{t,i}$$

for all $i = 1, \cdots, N$ and all $t = 1, \cdots, T$. The same holds for $b_i(t)^T(\mu - \hat{\mu}(t))$. Hence, with probability at least $1 - \delta$,

$$|b_i(t)^T(\hat{\mu}(t) - \mu)| \leq 2Rs_{t,i}\sqrt{\log(\frac{2Nt}{\delta})} + s_{t,i}, \tag{2.8}$$

for all i and all t. Hence, (2.3) holds when $\alpha = 2R\sqrt{\log(\frac{2NT}{\delta})} + 1$. However, in Algorithm 1 which chooses arms based on past rewards, each w_τ is correlated with the whole sequence $\{\eta_{a(1)}(1), \cdots, \eta_{a(t-2)}(t-2)\}$ through $B(t)$, so $\eta_{a(\tau)}(\tau)$'s are not independent given w_τ's. Therefore, the first equality in (2.7) does not hold and (2.8) does not hold either.

Abbasi-Yadkori et al. (2011) derived an upper bound for $b_i(t)^T\mu$ without requiring the assumption that $\eta_{a(\tau)}(\tau)$'s are independent. Hereinafter, we define $||x||_A := \sqrt{x^T A x}$ for any d-dimensional vector x and any $d \times d$ matrix A. First by Cauchy-Schwarz inequality,

$$\begin{aligned}
\left|b_i(t)^T(\hat{\mu}(t) - \mu)\right| &= \left|b_i(t)^T B(t)^{-1}\Big\{\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau) r_{a(\tau)}(\tau) - B(t)\mu\Big\}\right| \\
&= \left|b_i(t)^T B(t)^{-1}\Big\{\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau)\eta_{a(\tau)}(\tau) - \mu\Big\}\right| \\
&\leq \left|b_i(t)^T B(t)^{-1}\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau)\eta_{a(\tau)}(\tau)\right| \\
&+ \left|b_i(t)^T B(t)^{-1}\mu\right| \\
&\leq s_{t,i}\Big\{\Big|\Big|\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau)\eta_{a(\tau)}(\tau)\Big|\Big|_{B(t)^{-1}} + 1\Big\}.
\end{aligned} \tag{2.9}$$

Hence, if $(\alpha - 1)$ is a sharp upper bound of $\big|\big|\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau)\eta_{a(\tau)}(\tau)\big|\big|_{B(t)^{-1}}$ for all t, $U_i(t)$ is a sharp upper bound of $b_i(t)^T\mu$ for all i and all t.

We define $\mathcal{H}_{t-1}$ as the history until time $t-1$, i.e., $\mathcal{H}_{t-1} = \{a(\tau), r_{a(\tau)}(\tau), \{b_i(\tau)\}_{i=1}^N, \tau = 1, \cdots, t-1\}$, and the filtration $\mathcal{F}_{t-1}$ as the union of $\mathcal{H}_{t-1}$ and $\{a(t), b_{a(t)}(t)\}$. Then since $\big(\eta_{a(\tau)}(\tau)|\mathcal{F}_{\tau-1}\big)$ is R-sub-Gaussian, we have for any $\lambda \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\exp\Big(\frac{\eta_{a(\tau)}(\tau)}{R}\lambda^T b_{a(\tau)}(\tau) - \frac{1}{2}\lambda^T b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T \lambda\Big)\Big|\mathcal{F}_{\tau-1}\Big] \leq 1$$

Therefore for any $\lambda \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\exp\Big(\lambda^T \sum_{\tau=1}^{t-1} \frac{\eta_{a(\tau)}(\tau)}{R} b_{a(\tau)}(\tau) - \frac{1}{2}\lambda^T \sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T \lambda\Big)\Big] \leq 1. \tag{2.10}$$

From (2.10) we can apply the following lemma, which is a simplified version of the Corollary 4.3 of de la Peña et al. (2004).

Lemma 2.2.2. *Let $X_\tau \in \mathbb{R}^d$ and $z_\tau \in \mathbb{R}$ be some random variables, $\tau = 1, \cdots, t$. Suppose $\exists d \times d$ positive semi-definite matrix $A(t)$ such that for any $\lambda \in \mathbb{R}^d$,*

$$\mathbb{E}\Big[\exp\Big\{\lambda^T \sum_{\tau=1}^{t} X_\tau z_\tau - \frac{1}{2}\lambda^T A(t)\lambda\Big\}\Big] \leq 1. \tag{2.11}$$

Then for any $0 < \delta < 1$ and any positive definite matrix Q, with probability at least $1-\delta$,

$$\Big|\Big|\sum_{\tau=1}^{t} X_\tau z_\tau\Big|\Big|^2_{(Q+A(t))^{-1}} \leq \log\Big(\frac{det\big(Q+A(t)\big)/det\big(Q\big)}{\delta^2}\Big).$$

Proof. The proof of the lemma is simple. Since (2.11) holds for $\forall \lambda \in \mathbb{R}^d$, it also holds when λ is generated from the normal distribution $\mathcal{N}(0_d, Q^{-1})$. Also, the expectation of the left hand side of (2.11) over the distribution of λ is smaller or equal to 1 as well. Using this fact and Markov's inequality, the bound is derived. □

Taking $X_\tau = b_{a(\tau)}(\tau)$, $z_\tau = \frac{\eta_{a(\tau)}(\tau)}{R}$, and $A(t) = \sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau) b_{a(\tau)}(\tau)^T$, (2.11) is satisfied due to (2.10). By taking $Q = I_d$ and using the lemma, we have with probability at least $1 - \frac{\delta}{t^2}$,

$$\Big|\Big|\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau)\eta_{a(\tau)}(\tau)\Big|\Big|_{B(t)^{-1}} \leq R\sqrt{\log\Big(\frac{det(B(t))}{(\delta/t^2)^2}\Big)} \leq R\sqrt{3d\log\big(\frac{t}{\delta}\big)}. \tag{2.12}$$

The last inequality is due to the determinant-trace inequality,

$$det(B(t)) \leq \Big(\frac{trace(B(t))}{d}\Big)^d \leq t^d.$$

Therefore, (2.3) holds for all i and t by setting $\alpha = R\sqrt{3d\log\big(\frac{T}{\delta}\big)} + 1$ and without requiring that $\eta_{a(\tau)}(\tau)$'s are independent over time.

Plugging the value of α in (2.6), we obtain the high-probability upper bound of the regret of Algorithm 1. With probability at least $1 - \delta$,

$$\begin{aligned} R(T) &\leq 2\Big(R\sqrt{3d\log\big(T/\delta\big)} + 1\Big)\sqrt{2dT\log(1 + T/d)} \\ &= O\Big(d\sqrt{T\log\big(T/\delta\big)\log(1 + T/d)}\Big). \end{aligned} \tag{2.13}$$

The bound (2.13) matches the lower bound $\Omega(d\sqrt{T})$ for inifinite N by a factor of $\log(T)$. When N is finite, (2.13) is slightly higher than the lower bound $\Omega(\sqrt{dT})$ by a factor of $\sqrt{d}\log(T)$.

2.2.2 Thompson sampling (TS) algorithm

Thompson sampling (Thompson, 1933) has been widely used as a simple heuristic based on Bayesian ideas. Convergence properties were first derived in Wyatt (1997) for the 2-arm case with binary rewards and the multi-arm case with continuous rewards but with constant mean for each arm. Agrawal and Goyal (2013) was the

first to propose and analyze the Thompson sampling algorithm for linear contextual MABs (Algorithm 2) .

Algorithm 2 TS algorithm (Agrawal and Goyal, 2013)

Set $v = R\sqrt{d\log(T/\delta)}$, $B = I_d$, $y = 0_d$.
for $t = 1, \cdots, T$ **do**
    Compute $\hat{\mu}(t) = B^{-1}y$.
    Sample $\tilde{\mu}(t)$ from distribution $\mathcal{N}(\hat{\mu}(t), v^2B^{-1})$.
    Pull arm $a(t) = \underset{1\leq i\leq N}{\text{argmax}}\ b_i(t)^T\tilde{\mu}(t)$ and get reward $r_{a(t)}(t)$.
    Update B and y:
    $B \leftarrow B + b_{a(t)}(t)b_{a(t)}(t)^T$, $y \leftarrow y + b_{a(t)}(t)r_{a(t)}(t)$.
end for

The heuristic of the algorithm is to randomly pull the arm according to the posterior probability that it is the optimal arm. This can be done by sampling $\tilde{\mu}(t)$ from the posterior distribution of μ at time t, and pulling the arm $a(t) = \underset{1\leq i\leq N}{\text{argmax}}\ b_i(t)^T\tilde{\mu}(t)$. The posterior distribution $\mathcal{N}(\hat{\mu}(t), v^2B(t)^{-1})$ is easily derived by assuming a gaussian prior $\mathcal{N}(0_d, v^2I_d)$ on μ and that $r_i(t)$ given μ follows a gaussian distribution $\mathcal{N}(b_i(t)^T\mu, v^2)$.

Meanwhile, we notice that under $\tilde{\mu}(t) \sim \mathcal{N}(\hat{\mu}(t), v^2B(t)^{-1})$, $b_i(t)^T\tilde{\mu}(t)$ follows $\mathcal{N}(b_i(t)^T\hat{\mu}(t), v^2s_{t,i}^2)$. If we set v to be a similar value to α in Section 2.2.1, Algorithm 2 looks similar to Algorithm 1. The only difference is that Algorithm 1 makes decisions based on the upper confidence bound $U_i(t) = b_i(t)^T\hat{\mu}(t) + \alpha s_{t,i}$, while Algorithm 2 makes decisions based on random $b_i(t)^T\tilde{\mu}(t)$ which has mean $b_i(t)^T\hat{\mu}(t)$ and standard deviation $vs_{t,i}$.

Agrawal and Goyal (2013) derived the high-probability upper

bound of $R(T)$ for the TS algorithm. This bound does not require the Bayesian framework nor the gaussian assumption for the rewards. Under (2.1) and (2.2), it can be shown that with probability greater than $1-\delta$,

$$R(T) \leq O(d^{\frac{3}{2}}\sqrt{T\log(Td)\log(T/\delta)}(\sqrt{\log(1+T/d)}+\sqrt{\log(1/\delta)}\ )). \tag{2.14}$$

The regret analysis of the TS algorithm requires additional work due to the randomness of $\tilde{\mu}(t)$. The following is a brief outline of the proof of Agrawal and Goyal (2013).

(a) A high-probability bound of $|b_i(t)^T(\hat{\mu}(t)-\mu)|$ is derived. This is just a direct application of (2.9) and (2.12) from Abbasi-Yadkori et al. (2011). Let $E^{\hat{\mu}}(t)$ be the event,

$$E^{\hat{\mu}}(t) = \big\{\forall i : |b_i(t)^T(\hat{\mu}(t)-\mu)| \leq s_{t,i}\Big(R\sqrt{3d\log\frac{T}{\delta}}+1\Big)\big\}.$$

Then for all $\delta \in (0,1)$, for all $t \geq 1$, $\mathbb{P}(E^{\hat{\mu}}(t)) \geq 1-\frac{\delta}{t^2}$.

(b) A high-probability bound of $|b_i(t)^T(\tilde{\mu}(t)-\hat{\mu}(t))|$ is derived, using the conditional gaussian distribution of $\tilde{\mu}(t)$. Let $E^{\tilde{\mu}}(t)$ be the event,

$$E^{\tilde{\mu}}(t) = \{\forall i : |b_i(t)^T(\tilde{\mu}(t)-\hat{\mu}(t))| \leq s_{t,i}v\sqrt{4d\log(Td)}\}.$$

For all $\delta \in (0,1)$, for all $t \geq 1$, $\mathbb{P}(E^{\tilde{\mu}}(t)) \geq 1-\frac{1}{T^2}$.

(c) Arms at each time t are divided into two groups, saturated and unsaturated arms. The set $C(t)$ of saturated arms at time t is defined as follows:

$$C(t) = \{i : b_i(t)^T\mu + g(T)s_{t,i} < b_{a^*(t)}(t)^T\mu\},$$

where $g(T) = R\sqrt{3d\log\frac{T}{\delta}} + 1 + v\sqrt{4d\log(Td)}$. Note that the optimal arm is unsaturated. Note also that from (a) and (b), $b_i(t)^T\mu + g(T)s_{t,i}$ is an upper bound of $b_i(t)^T\tilde{\mu}(t)$. Hence by definition, the saturated arms are the arms that have quite accurate values of $b_i(t)^T\tilde{\mu}(t)$ so that their upper bound is lower than $b_{a^*(t)}(t)^T\mu$, enabling the algorithm to distinguish between them and the optimal arm.

(d) Define filtation $\mathcal{F}_{t-1} = \{\mathcal{H}_{t-1}, \{b_i(t)\}_{i=1}^N\}$. Given $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true, the probability of playing saturated arms is shown to be bounded by a function of the probability of playing unsaturated arms.

$$\mathbb{P}\big(a(t) \in C(t)|\mathcal{F}_{t-1}\big) \leq \frac{1}{p}\mathbb{P}\big(a(t) \notin C(t)|\mathcal{F}_{t-1}\big) + \frac{1}{pT^2},$$

where $p = \frac{1}{4e\sqrt{\pi}}$.

(e) From the definition of unsaturated arms and (d), the regret is bounded by a factor of $s_{t,a(t)}$ in expectation, which can be shown as follows. Let $\tilde{a}(t) = \underset{i \notin C(t)}{\operatorname{argmin}}\, s_{t,i}$. This value is determined by $\mathcal{F}_{t-1}$. Under both $E^{\hat{\mu}}(t)$ and $E^{\tilde{\mu}}(t)$,

$$\begin{aligned}
b_{a^*(t)}(t)^T\mu &= b_{a^*(t)}(t)^T\mu - b_{\tilde{a}(t)}(t)^T\mu + b_{\tilde{a}(t)}(t)^T\mu \\
&\leq g(T)s_{t,\tilde{a}(t)} + b_{\tilde{a}(t)}(t)^T\mu \\
&\leq g(T)s_{t,\tilde{a}(t)} + b_{\tilde{a}(t)}(t)^T\tilde{\mu}(t) + g(T)s_{t,\tilde{a}(t)} \\
&\leq 2g(T)s_{t,\tilde{a}(t)} + b_{a(t)}(t)^T\tilde{\mu}(t) \\
&\leq 2g(T)s_{t,\tilde{a}(t)} + b_{a(t)}(t)^T\mu + g(T)s_{t,a(t)} \\
\Rightarrow regret(t) &\leq 2g(T)s_{t,\tilde{a}(t)} + g(T)s_{t,a(t)},
\end{aligned}$$

where the first inequality follows from the definition of unsaturated arms, the second and fourth inequalities from $E^{\hat{\mu}}(t)$ and $E^{\tilde{\mu}}(t)$, and the third inequality from the action selection mechanism. Therefore, given $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ holds,

$$\begin{aligned}\mathbb{E}\big[regret(t)|\mathcal{F}_{t-1}\big] &\leq 2g(T)s_{t,\tilde{a}(t)} + g(T)\mathbb{E}\big[s_{t,a(t)}|\mathcal{F}_{t-1}\big] \\ &\quad + 1 - \mathbb{P}(E^{\tilde{\mu}}(t)|\mathcal{F}_{t-1}) \\ &\leq 2g(T)s_{t,\tilde{a}(t)} + g(T)\mathbb{E}\big[s_{t,a(t)}|\mathcal{F}_{t-1}\big] + \frac{1}{T^2}.\end{aligned} \tag{2.15}$$

Here,

$$\begin{aligned}s_{t,\tilde{a}(t)} &= s_{t,\tilde{a}(t)}\big\{\mathbb{P}(a(t) \in C(t)|\mathcal{F}_{t-1}) + \mathbb{P}(a(t) \notin C(t)|\mathcal{F}_{t-1})\big\} \\ &\leq s_{t,\tilde{a}(t)}\Big\{\frac{2}{p}\mathbb{P}(a(t) \notin C(t)|\mathcal{F}_{t-1}) + \frac{1}{pT^2}\Big\} \\ &= \frac{2}{p}\mathbb{E}\big(s_{t,\tilde{a}(t)}I\{a(t) \notin C(t)\}\big|\mathcal{F}_{t-1}\big) + \frac{s_{t,\tilde{a}(t)}}{pT^2} \\ &\leq \frac{2}{p}\mathbb{E}\big(s_{t,a(t)}I\{a(t) \notin C(t)\}\big|\mathcal{F}_{t-1}\big) + \frac{s_{t,\tilde{a}(t)}}{pT^2} \\ &\leq \frac{2}{p}\mathbb{E}\big(s_{t,a(t)}\big|\mathcal{F}_{t-1}\big) + \frac{1}{pT^2},\end{aligned}$$

where the first inequality is due to (d) and the second inequality is due to the definition of $\tilde{a}(t)$. Combining this result with (2.15), we have

$$\mathbb{E}\big[regret(t)|\mathcal{F}_{t-1}\big] \leq \frac{5g(T)}{p}\mathbb{E}\big(s_{t,a(t)}\big|\mathcal{F}_{t-1}\big) + \frac{3g(T)}{pT^2}.$$

(f) Due to (e), the sequence $\{M_t\}_{t=1}^T$ where $M_t = regret(t)I(E^{\hat{\mu}}(t)) - \frac{5g(T)}{p}s_{t,a(t)} - \frac{3g(T)}{pT^2}$ is a bounded super-martingale difference process with respect to the filtration $\{\mathcal{F}_t\}_{t=1}^T$. By applying Azuma-Hoeffding's inequality, we can bound $\sum_{t=1}^T M_t$ with

high probability. Then from (a) and (2.5), the regret bound is derived.

The bound (2.14) matches the bound (2.13) by a factor of $\sqrt{d}\sqrt{\log(T)}$, which is the price for randomness. On the other hand, the TS algorithm does not require the **for** loop in the UCB algorithm to compute the $s'_{t,i}s$ for each arm i.

2.3 Adversarial MAB

Linear contextual MABs reviewed in Section 2.2 rely on time-invariant, linear assumptions on the reward. In real world problems however, the reward distributions often change as time passes, according to the past actions or in a completely unexpected manner. Again in the news article recommendation example, the probability that a specific user clicks a specific article in the Featured tab can change over time, depending on the user's mood or whether the user has enough time to read the article at the moment he or she visits the homepage.

Adversarial contextual MABs constitute another big branch in the bandit literature. Unlike linear contextual MABs, they impose no distributional assumption on the reward $r_i(t)$ except that $|r_i(t)| < c$ for some $c > 0$. Hence, the distribution of $r_i(t)$ is allowed to change over time, and it can also change adaptively depending on $\mathcal{H}_{t-1}$. In fact, we assume that an unknown adversary controls the value of $r_i(t)$ in a way that hampers the learner. In this more relaxed setting though, it is hard to achieve low $regret(t)$ with respect to the best choice $r_{a^*(t)}(t)$.

In the adversarial setting, the goal is to minimize the regret with respect to the best policy among a finite set of predefined policies. Let K be the number of predefined policies. The j-th policy can depend on time t and is represented by a N-dimensional probability vector, $\xi^j(t) \in \mathbb{R}^N$ $(j = 1, \cdots, K)$, where the i-th element $\xi_i^j(t)$ denotes the probability of pulling the i-th arm at time t by the j-th policy. Then, the expectation of the reward obtained by following the j-th policy at time t is $y_j(t) := \xi^j(t)^T r(t)$, where $r(t) = [r_1(t), \cdots, r_N(t)]^T$ is the whole reward vector. Define $G_j := \sum_{t=1}^{T} y_j(t)$, which is the sum of expected rewards obtained by following the j-th policy for all T steps. A new notion of regret is

$$R'(T) := \max_{1 \le j \le K} G_j - \sum_{t=1}^{T} r_{a(t)}(t).$$

2.3.1 EXP4.P algorithm

Beygelzimer et al. (2011) proposed the EXP4.P algorithm (Algorithm 3) for adversarial MABs. At each iteration, the algorithm updates the weight $w_j(t)$ of the j-th policy by a factor of $\exp\{\frac{p_{min}}{2} u_j(t)\}$, where it can be shown that $\sum_t u_j(t)$ is a high-probability upper confidence bound of $\{G_j - \sqrt{NT\log(K/\delta)}\}$ for every $j = 1, \cdots, K$. The value $u_j(t)$ is computed using an unbiased, inverse probability weighting (IPW) estimator $\hat{y}_j(t)$ of $y_j(t)$ and its variance estimator $\hat{v}_j(t)$. The weights are then normalized to sum up to 1 and the probability of pulling the i-th arm, $p_i(t)$, is computed as the weighted average of $\xi_i^j(t)'s$ under restriction that $p_i(t) \ge p_{min}$ for every i. This restriction is required to prevent the IPW estimator $\hat{y}_j(t)$ from blowing up. This restriction

Algorithm 3 EXP4.P algorithm (Beygelzimer et al., 2011)

Set $\delta \in (0,1)$, $p_{min} = \sqrt{\frac{\log K}{NT}}$, $w_j(1) = 1$ for $j = 1, \cdots, K$.
for $t = 1, \cdots, T$ **do**
  for $i = 1, \cdots, N$ **do**
    Compute $p_i(t) = (1 - Np_{min}) \sum_{j=1}^{K} \frac{w_j(t)\xi_i^j(t)}{\sum_{l=1}^{K} w_l(t)} + p_{min}$.
  end for
  Pull arm $a(t)$ with probability $\mathbb{P}(a(t) = i) = p_i(t)$.
  Get reward $r_{a(t)}(t)$.
  for $i = 1, \cdots, N$ **do**
    Compute $\hat{r}_i(t) = r_i(t) I(a(t) = i)/p_i(t)$.
  end for
  for $j = 1, \cdots, K$ **do**
    Compute $\hat{y}_j(t) = \xi^j(t)^T \hat{r}(t)$, $\hat{v}_j(t) = \sum_{i=1}^{N} \xi_i^j(t)/p_i(t)$.
    Update $w_j(t+1)$:

$$w_j(t+1) = w_j(t) e^{\left(\frac{p_{min}}{2}\left(\hat{y}_j(t) + \hat{v}_j(t)\sqrt{\frac{\log(K/\delta)}{NT}}\right)\right)}.$$

  end for
end for

however introduces extra exploration that might be unnecessary if a bounded estimator were available.

Beygelzimer et al. (2011) proved that for $\forall \delta \in (0,1)$, if $r_i(t) \in [0,1]$, $\log(K/\delta) \leq NT$, and the set of policies includes one which, on each round, selects each action with probability $\frac{1}{N}$, the following bound holds for Algorithm 3 with probability at least $1 - \delta$:

$$R'(T) \leq 6\sqrt{TN\log(K/\delta)}. \tag{2.16}$$

Neu (2015) proposed a biased but bounded estimator of $y_j(t)$ to remove extra exploration induced by p_{min} and improved the above regret bound by a constant factor, from $6\sqrt{TN\log(K/\delta)}$ to $2\sqrt{2}\sqrt{TN\log(K/\delta)}$.

Although the fact that the bound (2.16) and the bound of Neu (2015) are valid under no distributional assumption on the reward is attractive, these bounds increase with the number of actions N and particularly, with the number of policies K. For $R'(T)$ to be close to the more conservative $R(T)$, K should be as large as possible so as to contain the optimal policy which chooses $a^*(t)$ for every t, leading to larger regret bounds. Therefore, when a simple parametric or semiparametric assumption is not considered so farfetched, algorithms that exploit this structure can lead to higher rewards.

2.4 Semiparametric contextual MAB

Greenewald et al. (2017) and Krishnamurthy et al. (2018) considered a middle ground between simple linear contextual MABs and complex adversarial MABs: a semiparametric contextual MAB. Hereinafter, we define the filtration $\mathcal{F}_{t-1}$ as the union of the history $\mathcal{H}_{t-1}$ and the contexts at time t, i.e., $\mathcal{F}_{t-1} = \{\mathcal{H}_{t-1}, \{b_i(t)\}_{i=1}^N\}$ for $t = 1, \cdots, T$. Given $\mathcal{F}_{t-1}$, they assumed that the expectation of the reward $r_i(t)$ can be decomposed into a time-invariant, linear component depending on action $(b_i(t)^T\mu)$ and a nonparametric component depending on time and possibly on $\mathcal{F}_{t-1}$, but not on

the action ($\nu(t)$):

$$\mathbb{E}(r_i(t)|\mathcal{F}_{t-1}) = \nu(t) + b_i(t)^T\mu. \tag{2.17}$$

In (2.17), we do not impose any distributional assumption on $\nu(t)$ except that it is bounded, $|\nu(t)| \leq 1$. If $\nu(t) = 0$, the problem is just a linear contextual MAB problem, whereas if $\nu(t)$ depends on the action as well, the reward distribution is completely nonparametric and can be addressed by adversarial MAB algorithms.

Greenewald et al. (2017) provided an intuitive motivation for using this semiparametric model in mobile heath (mHealth) settings. One goal of mHealth is to encourage the user to walk by sending particular messages that either suggest to go for a walk or just simply move around. The different messages correspond to the arms and the reward is the step count measured during a short time right after the message was sent. In such setting, $\nu(t)$ is the baseline reward, i.e., the number of steps when no message is sent, which can change with time depending on the abrupt mood of the user, or in a way that depends on the messages sent in the past and the user's past responses. On the other hand, it is often reasonable to assume that the amount of variation in the step counts when a particular message is sent compared to when no message is sent follows a stationary, linear model with respect to the context information gathered by the device.

In the familiar news article recommendation example, $\nu(t)$ can represent the baseline tendency of the user visiting at time t to click any article in the Featured tab, regardless of the contents of the article. This baseline tendency can change in an unexpected manner, because different users visit at each time and even for

the same user, the clicking tendency can change according to the user's mood or schedule, both of which cannot be captured as contextual information. Still, it is reasonable to assume that given this baseline tendency, the probability that the user clicks the given article is linear with respect to context information of the article and the user.

Under (2.17), we note that the optimal action $a^*(t)$ at time t does not depend on $\nu(t)$ but only on the value of μ, and the regret does not depend on $\nu(t)$ either: $regret(t) = b_{a^*(t)}(t)^T\mu - b_{a(t)}(t)^T\mu$. However, $\nu(t)$ confounds the estimation of μ. The nature of the bandit problem renders the distinction of $\nu(t)$ from the linear part especially difficult because only one observation is allowed at each time t. Moreover, under the partially adversarial model (2.17), deterministic algorithms such as UCB algorithms turn out to be useless. This is because for deterministic algorithms, $a(t) \in \mathcal{F}_{t-1}$. Hence, if an adversary sets $\nu(t) \in \mathcal{F}_{t-1}$ to be $\nu(t) = -b_{a(t)}(t)^T\mu$, the observed reward is $r_{a(t)}(t) = \eta_{a(t)}(t)$ for all $t = 1, \cdots, T$, and the algorithm cannot learn μ. Therefore, we should capitalize on the randomness of action choice.

Besides (2.17), we make the usual assumption that given $\mathcal{F}_{t-1}$, the error $\eta_i(t) := r_i(t) - \mathbb{E}(r_i(t)|\mathcal{F}_{t-1})$ is R-sub-Gaussian for some $R > 0$, i.e., for every $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda\eta_i(t))|\mathcal{F}_{t-1}] \leq \exp(\frac{\lambda^2 R^2}{2}). \tag{2.18}$$

Also without loss of generality, we assume

$$||b_i(t)||_2 \leq 1, \ ||\mu||_2 \leq 1, \ |\nu(t)| \leq 1. \tag{2.19}$$

2.4.1 Action-centered TS algorithm

Greenewald et al. (2017) proposed the action-centered TS algorithm (Algorithm 4) for the new reward model (2.17). In their settings, they assumed that the first action is the base action, of which the context vector is $b_1(t) = 0_d$ for all t. Hence, the expected reward of the base action is $\nu(t)$, which can vary with time, and also in a way that depends on the past. In the aforementioned mHealth settings, the non-base actions are the different messages encouraging to walk while the base action is "not sending any message". Greenewald et al. (2017) followed the basic framework of the randomized, TS algorithm but in 2 stages. In the first stage, the learner selects one action among the non-base actions in the same way as in TS algorithm using random $\tilde{\mu}(t)$. Let this action be $\bar{a}(t)$. In the second stage, the learner chooses once more between $\bar{a}(t)$ and the base action using the distribution of $\tilde{\mu}(t)$. This finally chosen action is set as $a(t)$ and only this action is actually taken. In the second stage, we can compute the probability of choosing $\bar{a}(t)$ over the base action as follows using the gaussian distribution of $\tilde{\mu}(t)$,

$$\begin{aligned}
\mathbb{P}(a(t) = \bar{a}(t)|\mathcal{F}_{t-1}, \bar{a}(t)) &= \mathbb{P}(b_{\bar{a}(t)}(t)^T\tilde{\mu}(t) > b_1(t)^T\tilde{\mu}(t)|\mathcal{F}_{t-1}, \bar{a}(t)) \\
&= \mathbb{P}(b_{\bar{a}(t)}(t)^T\tilde{\mu}(t) > 0|\mathcal{F}_{t-1}, \bar{a}(t)) \\
&= 1 - \psi\Big(\frac{-b_{\bar{a}(t)}(t)^T\hat{\mu}(t)}{vs_{t,\bar{a}(t)}(t)}\Big),
\end{aligned}$$

where $\psi(\cdot)$ is the CDF of the standard gaussian distribution.

Instead of choosing $a(t) = \bar{a}(t)$ with this exact probability however, Greenewald et al. (2017) constrained the probability of not choosing the base action to lie in a predefined set $[p_{min}, p_{max}] \subset$

Algorithm 4 Action-Centered TS (Greenewald et al., 2017)

Set $B = I_d$, $y = 0_d$, $v = R\sqrt{d\log(T/\delta)}$.
Choose p_{min}, $p_{max} \in (0, 1)$.
for $t = 1, 2, \cdots, T$ **do**
    Compute $\hat{\mu}(t) = B^{-1}y$.
    Sample $\tilde{\mu}(t)$ from distribution $\mathcal{N}(\hat{\mu}(t), v^2B^{-1})$.
    Compute $\bar{a}(t) := \underset{i\in\{2,\cdots,N\}}{\text{argmax}}\ b_i(t)^T\tilde{\mu}(t)$.
    Compute probability p_t of taking non-base action:

$$p_t = \max\Big(p_{min}, \min\big(\mathbb{P}_{\tilde{\mu}}(b_{\bar{a}(t)}(t)^T\tilde{\mu}(t) > 0|\mathcal{F}_{t-1}, \bar{a}(t)), p_{max}\big)\Big).$$

    Pull arm $a(t) = \bar{a}(t)$ with probability p_t, else pull $a(t) = 1$.
    Get reward $r_{a(t)}(t)$ and compute pseudo reward $\hat{r}_{\bar{a}(t)}(t)$.
    Update B and y:
    $B \leftarrow B + p_t(1 - p_t)b_{\bar{a}(t)}(t)b_{\bar{a}(t)}(t)^T, y \leftarrow y + b_{\bar{a}(t)}(t)\hat{r}_{\bar{a}(t)}(t)$.
end for

$[0, 1]$. This is to prevent the policy from converging to either "not sending any message at all", which can cause the user to disengage with the system, or "always sending a message", which can get the user be overwhelmed by the interventions. Hence, the algorithm selects $a(t) = \bar{a}(t)$ with probability

$$p_t = \max\Big(p_{min},\ \min\Big(1 - \psi\Big(\frac{-b_{\bar{a}(t)}(t)^T\hat{\mu}(t)}{vs_{t,\bar{a}(t)}(t)}\Big),\ p_{max}\Big)\Big).$$

Under this probability constraint, the definition of the optimal policy and $regret(t)$ changes accordingly. Let $\bar{a}^*(t) = \underset{2\le i\le N}{\text{argmax}}\ b_i(t)^T\mu$. Thus, $\bar{a}^*(t)$ is the optimal action among the non-base actions. Then the optimal policy chooses the action $a^*(t) = \bar{a}^*(t)$ with

probability $\pi^*(t) := p_{max} I(b_{\bar{a}^*(t)}(t)^T \mu > 0) + p_{min} I(b_{\bar{a}^*(t)}(t)^T \mu \leq 0)$ and $a^*(t) = 1$ with probability $1 - \pi^*(t)$.

To consistently estimate μ, Greenewald et al. (2017) defined a pseudo-reward,

$$\hat{r}_{\bar{a}(t)}(t) = \{I(a(t) = \bar{a}(t)) - p_t\} r_{a(t)}(t).$$

An important property of the pseudo-reward is that its expectation does not depend on $\nu(t)$. This can be shown as follows,

$$\begin{aligned}
\mathbb{E}(\hat{r}_{\bar{a}(t)}(t)|\mathcal{F}_{t-1}, \bar{a}(t)) &= p_t(1-p_t)\mathbb{E}(r_{\bar{a}(t)}(t)|\mathcal{F}_{t-1}, \bar{a}(t)) \\
&\quad + (1-p_t)(-p_t)\mathbb{E}(r_1(t)|\mathcal{F}_{t-1}, \bar{a}(t)) \\
&= p_t(1-p_t)\mathbb{E}(r_{\bar{a}(t)}(t) - r_1(t)|\mathcal{F}_{t-1}, \bar{a}(t)) \\
&= p_t(1-p_t) b_{\bar{a}(t)}(t)^T \mu.
\end{aligned}$$

Algorithm 4 uses $p_t(1-p_t)b_{\bar{a}(t)}(t)$ and $\hat{r}_{\bar{a}(t)}(t)$ instead of $b_{a(t)}(t)$ and $r_{a(t)}(t)$ to update B and y, and hence $\hat{\mu}(t)$. It can be shown that high probability upper bound of $R(T)$ for the action-centered TS algorithm matches that of the original TS algorithm for linear reward models, but by a constant factor $M = 1/\{p_{min}(1-p_{max})\}$. For $\forall \delta \in (0,1)$, with probability at least $1-\delta$,

$$R(T) \leq O\Big(M d^{\frac{3}{2}} \sqrt{T \log(Td) \log(T/\delta)} \big(\sqrt{\log(1+T/d)} + \sqrt{\log(1/\delta)}\big)\Big).$$

This factor M explodes in the case where we don't want to restrict action selection probabilities.

2.4.2 BOSE algorithm

Krishnamurthy et al. (2018) proposed the BOSE (Bandit Orthogonalized Semiparametric Estimation) algorithm for the semipara-

metric reward model (2.17). This algorithm takes a different approach from the algorithms using optimism principles. It uses an action elimination method adapted from Even-Dar et al. (2006). At each time t, an action i is eliminated if there exists another action j such that

$$\big(b_j(t) - b_i(t)\big)^T \hat{\mu}(t) > \omega \big|\big| b_i(t) - b_j(t) \big|\big|_{V_t^{-1}},$$

where ω is a predefined constant, $\hat{\mu}(t)$ is an estimate of μ, and V_t is a d-dimensional matrix. The algorithm then picks up one action randomly among the survivors according to a particular distribution.

For estimating μ, Krishnamurthy et al. (2018) used a centering trick on the context vectors $b_i(t)$'s to cancel out $\nu(t)$. They proposed the following estimator for μ:

$$\hat{\mu}(t) = \Big(\gamma I_d + \sum_{\tau=1}^{t-1} X_\tau X_\tau^T\Big)^{-1} \sum_{\tau=1}^{t-1} X_\tau r_{a(\tau)}(\tau), \tag{2.20}$$

where $X_\tau = b_{a(\tau)}(\tau) - \mathbb{E}(b_{a(\tau)}(\tau)|\mathcal{F}_{\tau-1})$ and $\gamma > 0$. Given $\mathcal{F}_{\tau-1}$, we see that $\mathbb{E}(X_\tau|\mathcal{F}_{\tau-1}) = 0_d$. Hence, $\{\sum_{\tau=1}^{t} X_\tau\}_{t=1}^{\infty}$ is a martingale adapted to filtration $\{\mathcal{F}_t\}_{t=1}^{\infty}$. Krishnamurthy et al. (2018) derived a $(1-\delta)$-probability upper bound for $b^T(\hat{\mu}(t) - \mu)$ using a new concentration inequality for self-normalized vector-valued martingales established by de la Peña et al. (2009) and de la Peña et al. (2004).

The BOSE algorithm does not require any constraint on the action choice probabilities but achieves a $O(d\sqrt{T}\log(T))$ regret bound. This bound matches the best known regret bound (2.13) for linear reward models. However, the action elimination step

requires $O(N^2)$ computations at each round. Also, the distribution used to select the action should satisfy a specific condition to guarantee the $O(d\sqrt{T}\log(T))$ regret bound. The authors however only showed the existence of a solution for this condition when $N > 2$. Furthermore, the bound of $b^T(\hat{\mu}(t) - \mu)$ is valid under $\gamma \geq 4d\log(9T) + 8\log(4T/\delta)$ when $N > 2$, which can overwhelm the denominator term of $\hat{\mu}(t)$ when t is small. For example, when $d = 35$ and $T = 1900000$ as in the news article recommendation example in Chapter 5, $\gamma \geq 2476.8$ if we take $\delta = 0.1$. When γ is set to be a tuning parameter, the BOSE algorithm requires in total 2 tuning parameters, including ω used in the action selection step.

Chapter 3

Proposed method

In this thesis, we propose a new algorithm for the semiparametric reward model (2.17) which improves on the results of Greenewald et al. (2017) while keeping the framework of the TS algorithm. Unlike Krishnamurthy et al. (2018), our method requires only $O(N)$ computations at each round and specifies an action selection distribution for every N. The proposed algorithm uses a new estimator $\hat{\mu}(t)$ for μ which enjoys a tighter high-probability upper bound than (2.20) without any big constant like γ. We prove that the high-probability upper bound of the regret $R(T)$ incurred by the proposed algorithm has the same order as the TS algorithm for linear reward models without the need to restrict action choice probabilities like in Greenewald et al. (2017).

3.1 Proposed algorithm

Besides (2.17), we make the same assumptions as in Section 2.4, (2.18) and (2.19). The proposed Algorithm 5 follows the framework

Algorithm 5 Proposed TS algorithm

Set $B = I_d$, $y = 0_d$, $v = (2R+6)\sqrt{6d\log(T/\delta)}$, $\delta \in (0,1)$.
for $t = 1, 2, \cdots, T$ **do**
  Compute $\hat{\mu}(t) = B^{-1}y$.
  Sample $\tilde{\mu}(t)$ from distribution $\mathcal{N}(\hat{\mu}(t), v^2 B^{-1})$.
  Pull arm $a(t) = \underset{1\le i\le N}{\operatorname{argmax}}\, b_i(t)^T \tilde{\mu}(t)$ and get reward $r_{a(t)}(t)$.
  for $i = 1, \cdots, N$ **do**
    Compute $\pi_i(t) = \mathbb{P}(a(t) = i|\mathcal{F}_{t-1})$.
  end for
  Update B and y:
  $B \leftarrow B + \big(b_{a(t)}(t) - \bar{b}(t)\big)\big(b_{a(t)}(t) - \bar{b}(t)\big)^T$,
  $B \leftarrow B + \sum_{i=1}^{N} \pi_i(t)\big(b_i(t) - \bar{b}(t)\big)\big(b_i(t) - \bar{b}(t)\big)^T$,
  $y \leftarrow y + 2\big(b_{a(t)}(t) - \bar{b}(t)\big) r_{a(t)}(t)$.
end for

of the TS algorithm (Algorithm 2) with 2 major modifications: the mean and variance of $\tilde{\mu}(t)$. First, we propose a new estimator $\hat{\mu}(t)$ of μ for the mean of $\tilde{\mu}(t)$:

$$\hat{\mu}(t) = \Big(I_d + \sum_{\tau=1}^{t-1} X_\tau X_\tau^T + \sum_{\tau=1}^{t-1} \mathbb{E}(X_\tau X_\tau^T|\mathcal{F}_{\tau-1})\Big)^{-1} \sum_{\tau=1}^{t-1} 2X_\tau r_{a(\tau)}(\tau), \tag{3.1}$$

where $X_\tau = b_{a(\tau)}(\tau) - \mathbb{E}(b_{a(\tau)}(\tau)|\mathcal{F}_{\tau-1})$. Compared to (2.20), we note that the proposed estimator stabilizes the denominator using a new term $\sum_{\tau=1}^{t-1} \mathbb{E}(X_\tau X_\tau^T|\mathcal{F}_{\tau-1})$ instead of γI_d. Hereinafter, let $\bar{b}(\tau)$ denote $\mathbb{E}(b_{a(\tau)}(\tau)|\mathcal{F}_{\tau-1})$ for simplicity. This term can be

calculated as

$$\bar{b}(\tau) = \mathbb{E}\Big(\sum_{i=1}^{N} I(a(\tau) = i) b_i(\tau) \Big| \mathcal{F}_{\tau-1} \Big) = \sum_{i=1}^{N} \pi_i(\tau) b_i(\tau),$$

where $\pi_i(\tau) = \mathbb{P}(a(\tau) = i | \mathcal{F}_{\tau-1})$ is the probability of pulling the i-th arm at time τ, which is determined by the distribution of $\tilde{\mu}(\tau)$. Also, the covariance $\mathbb{E}(X_\tau X_\tau^T | \mathcal{F}_{\tau-1})$ can be computed as

$$\mathbb{E}(X_\tau X_\tau^T | \mathcal{F}_{\tau-1}) = \sum_{i=1}^{N} \pi_i(\tau)(b_i(\tau) - \bar{b}(\tau))(b_i(\tau) - \bar{b}(\tau))^T.$$

As for the variance of $\tilde{\mu}(t)$, we propose $v^2 B(t)^{-1}$, where $v = (2R + 6)\sqrt{6d\log(T/\delta)}$ and $B(t) = I_d + \sum_{\tau=1}^{t-1} \big(b_{a(\tau)}(\tau) - \bar{b}(\tau)\big)\big(b_{a(\tau)}(\tau) - \bar{b}(\tau)\big)^T + \sum_{\tau=1}^{t-1} \sum_{i=1}^{N} \pi_i(\tau)\big(b_i(\tau) - \bar{b}(\tau)\big)\big(b_i(\tau) - \bar{b}(\tau)\big)^T$.

In the following theorem, we establish a high-probability regret upper bound for the proposed algorithm.

Theorem 3.1.1. *Under (2.17), (2.18), and (2.19), the regret of Algorithm 5 is bounded as follows. For* $\forall \delta \in (0, 1)$*, with probability* $1 - \delta$*,*

$$R(T) \leq O\Big(d^{3/2}\sqrt{T}\sqrt{\log(Td)\log(T/\delta)}\big(\sqrt{\log(1 + T/d)} + \sqrt{\log(1/\delta)}\big)\Big).$$

This bound matches the bound (2.14) of the original TS algorithm for linear reward models.

3.2 Proof

The proof of Theorem 3.1.1 follows stages (a)-(f) in Section 2.2.2. given by Agrawal and Goyal (2013) with some modifications. The main contribution of this thesis is a new theorem for stage (a) to bound $|(b_i(t) - \bar{b}(t))^T(\hat{\mu}(t) - \mu)|$ with the new estimator (3.1).

3.2.1 Stage (a)

We establish a tight high-probability upper bound of $|(b_i(t) - \bar{b}(t))^T(\hat{\mu}(t) - \mu)|$ in the following Theorem 3.2.1.

Theorem 3.2.1. *Let the event $E^{\hat{\mu}}(t)$ be defined as follows:*

$$E^{\hat{\mu}}(t) = \big\{\forall i : |\big(b_i(t) - \bar{b}(t)\big)^T(\hat{\mu}(t) - \mu)| \leq l(T)s^c_{t,i}\big\},$$

where $l(T) = (2R+6)\sqrt{d\log(6T^3/\delta)} + 1$ and $s^c_{t,i} = \sqrt{\big(b_i(t) - \bar{b}(t)\big)^T B(t)^{-1}\big(b_i(t) - \bar{b}(t)\big)}$. Then for all $t \geq 1$, for any $0 < \delta < 1$, $\mathbb{P}(E^{\hat{\mu}}(t)) \geq 1 - \frac{\delta}{t^2}$. It is trivial that $\mathbb{P}(E^{\hat{\mu}}(1)) = 1$.

Proof. By decomposition,

$$\begin{aligned}
\hat{\mu}(t) - \mu &= B(t)^{-1}\sum_{\tau=1}^{t-1} 2X_\tau r_{a(\tau)}(\tau) - \mu \\
&= B(t)^{-1}\big\{\sum_{\tau=1}^{t-1} 2X_\tau \eta_{a(\tau)}(\tau) + \sum_{\tau=1}^{t-1} 2X_\tau\big(\nu(\tau) + \bar{b}(\tau)^T\mu\big) \\
&\quad - \mu + \sum_{\tau=1}^{t-1} D(\tau)\mu\big\},
\end{aligned}$$

where $D(\tau) = X_\tau X_\tau^T - \mathbb{E}\big(X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big)$. Let $b^c_i(t) := b_i(t) - \bar{b}(t)$.

Then we have,

$$\begin{aligned}
\big|b_i^c(t)^T(\hat{\mu}(t)-\mu)\big| &= \big|b_i^c(t)^T B(t)^{-1}\big\{\sum_{\tau=1}^{t-1} 2X_\tau \eta_{a(\tau)}(\tau) \\
&\quad + \sum_{\tau=1}^{t-1} 2X_\tau\big(\nu(\tau)+\bar{b}(\tau)^T\mu\big) - \mu + \sum_{\tau=1}^{t-1} D(\tau)\mu\big\}\big| \\
&\le 2\big|b_i^c(t)^T B(t)^{-1}\sum_{\tau=1}^{t-1} X_\tau \eta_{a(\tau)}(\tau)\big| \\
&\quad + 2\big|b_i^c(t)^T B(t)^{-1}\sum_{\tau=1}^{t-1} X_\tau\big(\nu(\tau)+\bar{b}(\tau)^T\mu\big)\big| \\
&\quad + \big|b_i^c(t)^T B(t)^{-1}\mu\big| + \big|b_i^c(t)^T B(t)^{-1}\sum_{\tau=1}^{t-1} D(\tau)\mu\big| \\
&\le s_{t,i}^c \times \big\{2C_1 + 2C_2 + C_3 + C_4\big\},
\end{aligned}$$

where

$$\begin{aligned}
C_1 &= \sqrt{\Big(\sum_{\tau=1}^{t-1} X_\tau \eta_{a(\tau)}(\tau)\Big)^T B(t)^{-1}\Big(\sum_{\tau=1}^{t-1} X_\tau \eta_{a(\tau)}(\tau)\Big)}, \\
C_2 &= \sqrt{\Big(\sum_{\tau=1}^{t-1} X_\tau\big(\nu(\tau)+\bar{b}(\tau)^T\mu\big)\Big)^T B(t)^{-1}\Big(\sum_{\tau=1}^{t-1} X_\tau\big(\nu(\tau)+\bar{b}(\tau)^T\mu\big)\Big)}, \\
C_3 &= \sqrt{\Big(\sum_{\tau=1}^{t-1} D(\tau)\mu\Big)^T B(t)^{-1}\Big(\sum_{\tau=1}^{t-1} D(\tau)\mu\Big)}, \\
C_4 &= \sqrt{\mu^T B(t)^{-1}\mu}.
\end{aligned}$$

The second inequality is due to Cauchy-Schwarz inequality. First, $C_4 \le 1$. Now we will bound C_1, C_2, and C_3. First, the term C_1 is a familiar term, which we can bound using the R-sub-Gaussianity of $\eta_{a(\tau)}(\tau)$ and the technique of Abbasi-Yadkori et al. (2011). Since $\eta_{a(\tau)}(\tau)$ is R-sub-gaussian given $\mathcal{F}_{\tau-1}$ and $a(\tau)$, we have for any

$\lambda \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\exp\Big(\frac{\eta_{a(\tau)}(\tau)}{R}\lambda^T X_\tau - \frac{1}{2}\lambda^T X_\tau X_\tau^T \lambda\Big)\Big|\mathcal{F}_{\tau-1}, a(\tau)\Big] \leq 1$$
$$\Rightarrow \mathbb{E}\Big[\exp\Big(\lambda^T \sum_{\tau=1}^{t-1} \frac{\eta_{a(\tau)}(\tau)}{R} X_\tau - \frac{1}{2}\lambda^T \sum_{\tau=1}^{t-1} X_\tau X_\tau^T \lambda\Big)\Big] \leq 1. \quad (3.2)$$

Taking $c_\tau = \frac{1}{R}\eta_{a(\tau)}(\tau)$, $Q = I_d + \sum_{\tau=1}^{t-1} \mathbb{E}\big[X_\tau X_\tau^T | \mathcal{F}_{\tau-1}\big]$, and $A(t) = \sum_{\tau=1}^{t-1} X_\tau X_\tau^T$, we see that (3.2) corresponds to condition (2.11) of Lemma 2.2.2. Also,

$$C_1 = R\sqrt{\Big(\sum_{\tau=1}^{t-1} X_\tau c_\tau\Big)^T \big(Q + A(t)\big)^{-1} \Big(\sum_{\tau=1}^{t-1} X_\tau c_\tau\Big)}.$$

Therefore by Lemma 2.2.2, for any $0 < \delta < 1$, with probability at least $1 - \frac{\delta}{3t^2}$,

$$C_1 \leq R\sqrt{\log\Big(\frac{det(Q + A(t))/det(Q)}{(\delta/(3t^2))^2}\Big)}$$
$$\leq R\sqrt{\log\Big(\frac{det(Q + A(t))}{(\delta/(3t^2))^2}\Big)} = R\sqrt{\log\Big(\frac{det(B(t))}{(\delta/(3t^2))^2}\Big)}. \quad (3.3)$$

Now, we need to bound C_2 and C_3, which are new terms that arise due to the $\nu(\tau)$'s and the use of a new estimator (3.1). Although C_2 looks similar to C_1, the term $(\nu(\tau) + \bar{b}(\tau)^T \mu)$ is not sub-Gaussian, so we cannot use the technique of Abbasi-Yadkori et al. (2011) anymore. Instead, we have $\mathbb{E}[X_\tau | \mathcal{F}_{\tau-1}] = 0$. To bound a similar term to C_2, Krishnamurthy et al. (2018) proposed to use Lemma 7 of de la Peña et al. (2009) for vector-valued martingales to derive an inequality analogous to (3.2). We first present the lemma of de la Peña et al. (2009), which is derived from Lemma 3.2.2 (Bercu and Touati, 2008).

Lemma 3.2.2. *(Lemma 2.1 of Bercu and Touati, 2008) Let x be a square integrable random variable with mean 0 and variance $\sigma^2 > 0$. Then,*

$$\mathbb{E}\Big[\exp\Big(x - \frac{1}{2}x^2 - \frac{1}{2}\sigma^2\Big)\Big] \leq 1.$$

Lemma 3.2.3. *(Lemma 7 of de la Peña et al., 2009) Let $X_\tau \in \mathbb{R}^d$ be $\mathcal{F}_\tau$-measurable for some filtration $\{\mathcal{F}_\tau\}_{\tau=1}^t$, $\mathbb{E}\big[X_\tau|\mathcal{F}_{\tau-1}\big] = 0$, and $||X_\tau||_2 \leq B$ for some constant B, $\tau = 1, \cdots, t$. Let $c_\tau \in \mathbb{R}$ be $\mathcal{F}_\tau$-measurable, $|c_\tau| \leq 1$ and $X_\tau \perp c_\tau|\mathcal{F}_{\tau-1}$. Then for any $\lambda \in \mathbb{R}^d$,*

$$\mathbb{E}\Big[\exp\Big\{\lambda^T \sum_{\tau=1}^t X_\tau c_\tau - \frac{1}{2}\lambda^T\Big(\sum_{\tau=1}^t X_\tau X_\tau^T + \sum_{\tau=1}^t \mathbb{E}\big[X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda\Big\}\Big] \leq 1.$$

Proof. Taking $x = \lambda^T X_\tau c_\tau$, we have from Lemma 3.2.2,

$$\mathbb{E}\Big[\exp\Big\{\lambda^T X_\tau c_\tau - \frac{1}{2}\lambda^T\Big(c_\tau^2 X_\tau X_\tau^T + \mathbb{E}\big[c_\tau^2 X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda\Big\}\Big|\mathcal{F}_{\tau-1}\Big] \leq 1.$$

Since $c_\tau^2 \leq 1$ and $X_\tau X_\tau^T$ is positive semi-definite,

$$\mathbb{E}\Big[\exp\Big\{\lambda^T X_\tau c_\tau - \frac{1}{2}\lambda^T\Big(X_\tau X_\tau^T + \mathbb{E}\big[X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda\Big\}\Big|\mathcal{F}_{\tau-1}\Big] \leq 1.$$

□

Now, we can derive an inequality analogous to (3.2). Take $c_\tau = \Big(\frac{\nu(\tau)+\bar{b}(\tau)^T\mu}{2}\Big)$. Since $\mathbb{E}\big[X_\tau|\mathcal{F}_{\tau-1}\big] = 0$, $|c_\tau| \leq 1$, and $X_\tau \perp c_\tau|\mathcal{F}_{\tau-1}$, we can apply Lemma 3.2.3, i.e., for any $\lambda \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\exp\Big\{\lambda^T \sum_{\tau=1}^{t-1} X_\tau c_\tau - \frac{1}{2}\lambda^T\Big(\sum_{\tau=1}^{t-1} X_\tau X_\tau^T + \sum_{\tau=1}^{t-1} \mathbb{E}\big[X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda\Big\}\Big] \leq 1. \tag{3.4}$$

Taking $A(t) = \sum_{\tau=1}^{t-1} X_\tau X_\tau^T + \sum_{\tau=1}^{t-1} \mathbb{E}\big[X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]$ and $Q = I_d$, (3.4) corresponds to condition (2.11). Also,

$$C_2 = 2\sqrt{\Big(\sum_{\tau=1}^{t-1} X_\tau c_\tau\Big)^T \big(Q + A(t)\big)^{-1}\Big(\sum_{\tau=1}^{t-1} X_\tau c_\tau\Big)}.$$

Therefore by Lemma 2.2.2, for any $0 < \delta < 1$, with probability at least $1 - \frac{\delta}{3t^2}$,

$$C_2 \leq 2\sqrt{\log\Big(\frac{det(Q + A(t))}{(\delta/(3t^2))^2}\Big)} = 2\sqrt{\log\Big(\frac{det(B(t))}{(\delta/(3t^2))^2}\Big)}. \qquad (3.5)$$

The final step is to bound C_3. However, C_3 does not take the form $||\sum X_\tau c_\tau||_{B(t)^{-1}}$, so we require additional work. Let $Y_\tau = D(\tau)\mu$. Then $Y_\tau \in \mathbb{R}^d$ and $\mathbb{E}\big[Y_\tau|\mathcal{F}_{\tau-1}\big] = 0$. By Lemma 3.2.3, for any $\lambda \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\exp\Big\{\lambda^T \sum_{\tau=1}^{t-1} \frac{1}{\sqrt{2}} Y_\tau - \frac{1}{2}\lambda^T\Big(\frac{1}{2}\sum_{\tau=1}^{t-1} Y_\tau Y_\tau^T + \frac{1}{2}\sum_{\tau=1}^{t-1}\mathbb{E}\big[Y_\tau Y_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda\Big\}\Big] \leq 1.$$

Here,

$$\begin{aligned}
\lambda^T Y_\tau Y_\tau^T \lambda &= \lambda^T D(\tau)\mu\mu^T D(\tau)\lambda \\
&= \big\{\big(D(\tau)\lambda\big)^T\mu\big\}^2 \\
&\leq \mu^T\mu\big(D(\tau)\lambda\big)^T\big(D(\tau)\lambda\big) \quad (\because \text{Cauchy-Schwarz inequality}) \\
&\leq \big(D(\tau)\lambda\big)^T\big(D(\tau)\lambda\big) = \lambda^T D(\tau)^2\lambda, \qquad (3.6)
\end{aligned}$$

and

$$\lambda^T\mathbb{E}\big[Y_\tau Y_\tau^T|\mathcal{F}_{\tau-1}\big]\lambda \leq \lambda^T\mathbb{E}\big[D(\tau)^2|\mathcal{F}_{\tau-1}\big]\lambda. \qquad (3.7)$$

Let $L = X_\tau X_\tau^T$ and $K = \mathbb{E}\big[X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]$. Then,

$$\begin{aligned}
\lambda^T D(\tau)^2\lambda &= \lambda^T\big(L - K\big)^2\lambda \\
&= \lambda^T L^2\lambda + \lambda^T K^2\lambda + 2\lambda^T L(-K)\lambda \\
&\leq \lambda^T L^2\lambda + \lambda^T K^2\lambda + 2\sqrt{\lambda^T L^2\lambda\ \lambda^T K^2\lambda} \\
&\quad (\because \text{Cauchy-Schwarz inequality}) \\
&\leq 2\lambda^T L^2\lambda + 2\lambda^T K^2\lambda. \qquad (3.8)
\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{E}\big[D(\tau)^2|\mathcal{F}_{\tau-1}\big] &= \mathbb{E}\big[(L-K)^2|\mathcal{F}_{\tau-1}\big] \\ &= \mathbb{E}\big[L^2|\mathcal{F}_{\tau-1}\big] - \mathbb{E}\big[L|\mathcal{F}_{\tau-1}\big]K - K\mathbb{E}\big[L|\mathcal{F}_{\tau-1}\big] \\ &\quad + K^2 \\ &= \mathbb{E}\big[L^2|\mathcal{F}_{\tau-1}\big] - K^2 \quad (\because \mathbb{E}\big[L|\mathcal{F}_{\tau-1}\big] = K) \\ \Rightarrow \lambda^T\mathbb{E}\big[D(\tau)^2|\mathcal{F}_{\tau-1}\big]\lambda &\leq 2\lambda^T\mathbb{E}\big[D(\tau)^2|\mathcal{F}_{\tau-1}\big]\lambda \\ &= 2\lambda^T\mathbb{E}\big[L^2|\mathcal{F}_{\tau-1}\big]\lambda - 2\lambda^T K^2\lambda. \end{aligned} \tag{3.9}$$

Due to (3.6), (3.7), (3.8) and (3.9),

$$\begin{aligned}\lambda^T\Big(Y_\tau Y_\tau^T + \mathbb{E}\big[Y_\tau Y_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda &\leq 2\lambda^T\Big(L^2 + \mathbb{E}\big[L^2|\mathcal{F}_{\tau-1}\big]\Big)\lambda \\ &\leq 2\lambda^T\Big(X_\tau X_\tau^T + \mathbb{E}\big[X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda,\end{aligned}$$

where the last inequality is due to $L = X_\tau X_\tau^T$ and $X_\tau^T X_\tau \leq 1$. Therefore, for any $\lambda \in \mathbb{R}^d$,

$$\mathbb{E}\Big[\exp\Big\{\lambda^T\sum_{\tau=1}^{t-1}\frac{1}{\sqrt{2}}Y_\tau - \frac{1}{2}\lambda^T\Big(\sum_{\tau=1}^{t-1}X_\tau X_\tau^T + \sum_{\tau=1}^{t-1}\mathbb{E}\big[X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda\Big\}\Big]$$

$$\leq \mathbb{E}\Big[\exp\Big\{\lambda^T\sum_{\tau=1}^{t-1}\frac{1}{\sqrt{2}}Y_\tau - \frac{1}{2}\lambda^T\Big(\frac{1}{2}\sum_{\tau=1}^{t-1}Y_\tau Y_\tau^T + \frac{1}{2}\sum_{\tau=1}^{t-1}\mathbb{E}\big[Y_\tau Y_\tau^T|\mathcal{F}_{\tau-1}\big]\Big)\lambda\Big\}\Big]$$

$$\leq 1. \tag{3.10}$$

Taking $A(t) = \sum_{\tau=1}^{t-1}X_\tau X_\tau^T + \sum_{\tau=1}^{t-1}\mathbb{E}\big[X_\tau X_\tau^T|\mathcal{F}_{\tau-1}\big]$ and $Q = I_d$, (3.10) corresponds to condition (2.11). Also,

$$C_3 = 2\sqrt{\Big(\sum_{\tau=1}^{t-1}\frac{1}{\sqrt{2}}Y_\tau\Big)^T\big(Q + A(t)\big)^{-1}\Big(\sum_{\tau=1}^{t-1}\frac{1}{\sqrt{2}}Y_\tau\Big)}.$$

Therefore by Lemma 2.2.2, for any $0 < \delta < 1$, with probability at least $1 - \frac{\delta}{3t^2}$,

$$C_3 \leq 2\sqrt{\log\Big(\frac{det(Q + A(t))}{(\delta/(3t^2))^2}\Big)} = 2\sqrt{\log\Big(\frac{det(B(t))}{(\delta/(3t^2))^2}\Big)}. \qquad (3.11)$$

Due to the bounds (3.3), (3.5) and (3.11), we have for any $0 < \delta < 1$, with probability at least $1 - \frac{\delta}{t^2}$,

$$\begin{aligned}\big|b_i^c(t)^T(\hat{\mu}(t) - \mu)\big| &\leq s_{t,i}^c \times \big\{2C_1 + 2C_2 + C_3 + C_4\big\} \\ &\leq s_{t,i}^c \times \Big\{(2R+6)\sqrt{\log\Big(\frac{det(B(t))}{(\delta/(3t^2))^2}\Big)} + 1\Big\},\end{aligned}$$

for all $i = 1, \cdots, N$. Due to the determinant-trace inequality,

$$det(B(t)) \leq \Big(\frac{trace(B(t))}{d}\Big)^d \leq \Big(\frac{d + 2(t-1)}{d}\Big)^d.$$

Hence, with probability at least $1 - \frac{\delta}{t^2}$, for all $i = 1, \cdots, N$,

$$\begin{aligned}\big|\big(b_i(t) - \bar{b}(t)\big)^T(\hat{\mu}(t) - \mu)\big| &\leq s_{t,i}^c \times \Big\{(2R+6)\sqrt{d\log\Big(\frac{3t^2 2t}{\delta}\Big)} + 1\Big\} \\ &\leq s_{t,i}^c \times \Big\{(2R+6)\sqrt{d\log\Big(\frac{6t^3}{\delta}\Big)} + 1\Big\} \\ &\leq l(T)s_{t,i}^c.\end{aligned}$$

□

3.2.2 Stage (b)

We next establish a high-probability upper bound of $|(b_i(t) - \bar{b}(t))^T(\tilde{\mu}(t) - \hat{\mu}(t))|$ in the following Proposition 3.2.5. The proof is a simple extension of Agrawal and Goyal (2013), which uses the following lemma for gaussian random variables.

Lemma 3.2.4. *(Abramowitz and Stegun, 1964) If $Z \sim \mathcal{N}(m, \sigma^2)$, for any $z \geq 1$,*

$$\frac{1}{2\sqrt{\pi}z}\exp\Big(-\frac{z^2}{2}\Big) \leq \mathbb{P}\big(|Z-m| > z\sigma\big) \leq \frac{1}{\sqrt{\pi}z}\exp\Big(-\frac{z^2}{2}\Big).$$

Proposition 3.2.5. *Let the event $E^{\tilde{\mu}}(t)$ be defined as follows:*

$$E^{\tilde{\mu}}(t) = \big\{\forall i : |\big(b_i(t) - \bar{b}(t)\big)^T(\tilde{\mu}(t) - \hat{\mu}(t))| \leq m(T)s^c_{t,i}\big\},$$

where $m(T) = v\sqrt{4d\log(Td)}$. Then for all $t \geq 0$, $\mathbb{P}(E^{\tilde{\mu}}(t)|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{T^2}$.

Proof. Note that given $\mathcal{F}_{t-1}$, the values of $\big(b_i(t) - \bar{b}(t)\big)$, $B(t)$ and $\hat{\mu}(t)$ are fixed. Then,

$$\begin{aligned}
|b_i^c(t)^T(\tilde{\mu}(t) - \hat{\mu}(t))| &= |b_i^c(t)^T v B(t)^{-1/2}\frac{1}{v}B(t)^{1/2}(\tilde{\mu}(t) - \hat{\mu}(t))| \\
&\leq v\sqrt{b_i^c(t)^T B(t)^{-1} b_i^c(t)}\Big|\Big|\frac{1}{v}B(t)^{1/2}(\tilde{\mu}(t) - \hat{\mu}(t))\Big|\Big|_2 \\
&= v s^c_{t,i}\Big|\Big|\frac{1}{v}B(t)^{1/2}(\tilde{\mu}(t) - \hat{\mu}(t))\Big|\Big|_2 \\
&= v s^c_{t,i}\sqrt{\sum_{j=1}^{d}||Z_j(t)||_2^2},
\end{aligned}$$

where $Z_j(t)|\mathcal{F}_{t-1} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and the first inequality is due to Cauchy-Schwarz inequality. Due to Lemma 3.2.4, for fixed j and $z \geq 1$,

$$\mathbb{P}\big(|Z_j(t)| > z \mid \mathcal{F}_{t-1}\big) \leq \frac{1}{\sqrt{\pi}z}\exp\Big(-\frac{\mathrm{z}^2}{2}\Big) \leq \exp\Big(-\frac{\mathrm{z}^2}{2}\Big).$$

Setting $\exp\big(-\mathrm{z}^2/2\big) = \frac{1}{dT^2}$, we have $z = \sqrt{2\log(dT^2)} \leq \sqrt{2\log(d^2T^2)}$ $= \sqrt{4\log(dT)}$. Hence,

$$\begin{aligned}
\mathbb{P}\big(|Z_j(t)| > \sqrt{4\log(dT)} \mid \mathcal{F}_{t-1}\big) &\leq \frac{1}{dT^2} \\
\Rightarrow \mathbb{P}\big(\forall j : |Z_j(t)| > \sqrt{4\log(dT)} \mid \mathcal{F}_{t-1}\big) &\leq \frac{1}{T^2}.
\end{aligned}$$

Thus, with probability at least $1 - \frac{1}{T^2}$, for all $i = 1, \cdots, N$,

$$|\big(b_i(t) - \bar{b}(t)\big)^T(\tilde{\mu}(t) - \hat{\mu}(t))| \leq v s_{t,i}^c \sqrt{4d\log(dT)} = m(T) s_{t,i}^c.$$

□

3.2.3 Stage (c)

Before proceeding, we divide the arms at each time into two groups: saturated and unsaturated arms. Let $g(T) = m(T) + l(T)$. An arm i is saturated at time t if

$$\big(b_i(t) - \bar{b}(t)\big)^T \mu + g(T) s_{t,i}^c < \big(b_{a^*(t)}(t) - \bar{b}(t)\big)^T \mu,$$

and unsaturated otherwise. Note that the optimal arm $a^*(t)$ is unsaturated. Although $\bar{b}(t)^T \mu$ can be canceled out in both sides, the definition of saturation is slightly different from Section 2.2.2 because we replaced $s_{t,i}$ with $s_{t,i}^c$.

3.2.4 Stage (d)

Next, we show in Proposition 3.2.6 that the probability of playing saturated arms is bounded by a function of the probability of playing unsaturated arms. The proof is a simple extension of Agrawal and Goyal (2013).

Proposition 3.2.6. *Let $C(t)$ be the set of saturated arms at time t, i.e., $C(t) = \{i : \big(b_i(t) - \bar{b}(t)\big)^T \mu + g(T) s_{t,i}^c < \big(b_{a^*(t)}(t) - \bar{b}(t)\big)^T \mu\}$. Given any filtration $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,*

$$\mathbb{P}\big(a(t) \in C(t) | \mathcal{F}_{t-1}\big) \leq \frac{1}{p} \mathbb{P}\big(a(t) \notin C(t) | \mathcal{F}_{t-1}\big) + \frac{1}{pT^2},$$

where $p = \frac{1}{4e\sqrt{2}\sqrt{\pi}}$.

Proof. Since the algorithm pulls the arm $\underset{i}{\text{argmax}}\{b_i(t)^T\tilde{\mu}(t)\}$, if $b_{a^*(t)}(t)^T\tilde{\mu}(t) > b_j(t)^T\tilde{\mu}(t)$ for every $j \in C(t)$, then $a(t) \notin C(t)$. Hence,

$$\begin{aligned}\mathbb{P}\big(a(t) \notin C(t)|\mathcal{F}_{t-1}\big) &\geq \mathbb{P}\big(b_{a^*(t)}(t)^T\tilde{\mu}(t) > b_j(t)^T\tilde{\mu}(t),\ \forall j \in C(t)|\mathcal{F}_{t-1}\big) \\ &= \mathbb{P}\big(b^c_{a^*(t)}(t)^T\tilde{\mu}(t) > b^c_j(t)^T\tilde{\mu}(t),\ \forall j \in C(t)|\mathcal{F}_{t-1}\big).\end{aligned} \tag{3.12}$$

If $E^{\tilde{\mu}}(t)$ is additionally true, for $\forall j \in C(t)$,

$$\begin{aligned}b^c_j(t)^T\tilde{\mu}(t) &\leq b^c_j(t)^T\mu + g(T)s^c_{t,j} \quad (\because E^{\hat{\mu}}(t)\ \&\ E^{\tilde{\mu}}(t)) \\ &\leq b^c_{a^*(t)}(t)^T\mu. \quad (\because \text{definition of } C(t))\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{P}\big(b^c_{a^*(t)}(t)^T\tilde{\mu}(t) > b^c_j(t)^T\tilde{\mu}(t),\ \forall j \in C(t)|\mathcal{F}_{t-1}\big) + \Big(1 - \mathbb{P}\big(E^{\tilde{\mu}}(t)|\mathcal{F}_{t-1}\big)\Big) \\ &\qquad \geq \mathbb{P}\big(b^c_{a^*(t)}(t)^T\tilde{\mu}(t) > b^c_{a^*(t)}(t)^T\mu|\mathcal{F}_{t-1}\big).\end{aligned} \tag{3.13}$$

Given $E^{\hat{\mu}}(t)$, $|b^c_{a^*(t)}(t)^T(\hat{\mu}(t) - \mu)| \leq l(T)s^c_{t,a^*(t)}$. Thus by Lemma 3.2.4,

$$\begin{aligned}(3.13) &= \mathbb{P}\Big(\ \frac{b^c_{a^*(t)}(t)^T(\tilde{\mu}(t) - \hat{\mu}(t))}{vs^c_{t,a^*(t)}} > \frac{b^c_{a^*(t)}(t)^T(\mu - \hat{\mu}(t))}{vs^c_{t,a^*(t)}}\Big|\mathcal{F}_{t-1}\Big) \\ &\geq \mathbb{P}\Big(\ Z(t) > \frac{l(T)}{v}\Big|\mathcal{F}_{t-1}\Big) \\ &\geq \frac{1}{4\sqrt{\pi}z}\exp\Big(-\frac{z^2}{2}\Big) \geq p,\end{aligned} \tag{3.14}$$

where $Z(t)|\mathcal{F}_{t-1} \sim \mathcal{N}(0,1)$ and $z = l(T)/v$. Therefore, due to (3.12), (3.13), (3.14) and Proposition 3.2.5,

$$\begin{aligned}&\mathbb{P}\big(a(t) \notin C(t)|\mathcal{F}_{t-1}\big) \geq p - \frac{1}{T^2}. \\ \Rightarrow\ &\frac{\mathbb{P}\big(a(t) \in C(t)|\mathcal{F}_{t-1}\big)}{\mathbb{P}\big(a(t) \notin C(t)|\mathcal{F}_{t-1}\big) + \frac{1}{T^2}} \leq \frac{1}{p}.\end{aligned}$$

□

3.2.5 Stage (e)

Next in Proposition 3.2.7, we use Proposition 3.2.6 and the definition of unsaturated arms to show that the regret can be bounded by a factor of $s^c_{t,a(t)}$ in expectation.

Proposition 3.2.7. *Given any filtration $\mathcal{F}_{t-1}$ such that $E^{\hat{\mu}}(t)$ is true,*

$$\mathbb{E}\big[regret(t)|\mathcal{F}_{t-1}\big] \leq \frac{5g(T)}{p}\mathbb{E}\big[s^c_{t,a(t)}|\mathcal{F}_{t-1}\big] + \frac{3g(T)}{pT^2}.$$

Proof. We simply replace $b_i(t)$ and $s_{t,i}$ in stage (e) of Section 2.2.2 with their centered versions, $b^c_i(t)$ and $s^c_{t,i}$. □

3.2.6 Stage (f)

Let $M_t = regret(t)I(E^{\hat{\mu}}(t)) - \frac{5g(T)}{p}s^c_{t,a(t)} - \frac{3g(T)}{pT^2}$. Then $|M_t|$ is bounded by $\frac{9g(T)}{p}$. Also, due to Proposition 3.2.7, $\{M_t\}_{t=1}^T$ is a bounded super-martingale difference process with respect to the filtration $\{\mathcal{F}_t\}_{t=1}^T$. Hence by Azuma-Hoeffding's inequality, for any $a \geq 0$,

$$\mathbb{P}\big(\sum_{t=1}^{T} M_t \geq a\big) \leq \exp\Big(-\frac{a^2}{2\sum_{t=1}^T c_t^2}\Big),$$

where $c_t = \frac{9}{p}g(T)$. Setting $\exp\Big(-\frac{a^2}{2\sum_{t=1}^T c_t^2}\Big) = \frac{\delta}{2}$, we have $a = \frac{9}{p}g(T)\sqrt{2T\log\big(\frac{2}{\delta}\big)}$. Thus with probability at least $1-\frac{\delta}{2}$,

$$\sum_{t=1}^{T} regret(t)I(E^{\hat{\mu}}(t)) \leq \frac{5g(T)}{p}\sum_{t=1}^{T} s^c_{t,a(t)} + \frac{3g(T)}{pT} + \frac{9}{p}g(T)\sqrt{2Tlog\big(\frac{2}{\delta}\big)}. \tag{3.15}$$

In Proposition 3.2.8, we show that $\sum_{t=1}^T s^c_{t,a(t)} \leq \sqrt{2dT\log(1+T/d)}$. This is a simple modification of (2.5).

Proposition 3.2.8. $\sum_{t=1}^{T} s^c_{t,a(t)} \leq \sqrt{2dT\log(1+T/d)}$.

Proof. Take $X_t = b_{a(t)}(t) - \bar{b}(t)$, $Q = I_d$, and $A(t) = \sum_{\tau=1}^{t-1} X_\tau X_\tau^T$. Then by Jensen's inequality and Lemma 2.2.1,

$$\begin{aligned}
\sum_{t=1}^{T} s^c_{t,a(t)} &\leq \sqrt{T\sum_{t=1}^{T}\{s^c_{t,a(t)}\}^2} \quad (\because \text{ Jensen's inequality}) \\
&= \sqrt{T\sum_{t=1}^{T} X_t^T B(t)^{-1} X_t} \\
&\leq \sqrt{T\sum_{t=1}^{T} X_t^T \{Q + A(t)\}^{-1} X_t} \quad (\because B(t) \succ Q + A(t)) \\
&\leq \sqrt{2T\log\Big(\frac{det(Q + A(T+1))}{det(Q)}\Big)} \quad (\because \text{ Lemma 2.2.1}) \\
&\leq \sqrt{2dT\log\Big(1 + \frac{T}{d}\Big)}. \quad (\because \text{ determinant-trace inequality.})
\end{aligned}$$

□

Due to (3.15), Proposition 3.2.8 and the definitions of p and $g(T)$, we have with probability at least $1 - \frac{\delta}{2}$,
$\sum_{t=1}^{T} regret(t) I(E^{\hat{\mu}}(t)) \leq$

$$O\Big(d^{3/2}\sqrt{T}\sqrt{\log(Td)\log(T/\delta)}\big(\sqrt{\log(1+T/d)} + \sqrt{\log(1/\delta)}\big)\Big).$$

Since $E^{\hat{\mu}}(t)$ holds for all t with probability at least $1 - \frac{\delta}{2}$ (Theorem 3.2.1), $regret(t) I(E^{\hat{\mu}}(t)) = regret(t)$ for all t with probability at least $1 - \frac{\delta}{2}$. Hence, with probability at least $1 - \delta$,

$$R(T) \leq O\Big(d^{3/2}\sqrt{T}\sqrt{\log(Td)\log(T/\delta)}\big(\sqrt{\log(1+T/d)} + \sqrt{\log(1/\delta)}\big)\Big).$$

Chapter 4

Simulation study

We conducted simulation studies to evaluate the proposed algorithm, the original TS algorithm (Agrawal and Goyal, 2013), and the action-centered TS algorithm of Greenewald et al. (2017). We set the number of actions as $N = 6$ and the dimension of the context vectors as $d = 10$. We set the first action to be the base action, i.e., $b_1(t) = 0_d$ for all t, and formed the other context vectors as

$$b_{i,j}(t) = \begin{cases} z_{i,1}(t) & \text{if } j = 2(i-1) - 1 \\ z_{i,2}(t) & \text{if } j = 2(i-1) \\ 0 & \text{otherwise} \end{cases}$$

where $z_i(t) \in \mathbb{R}^2$ is generated i.i.d. and uniformly from the unit circle. We generated $\eta_i(t) \overset{i.i.d.}{\sim} \mathcal{N}(0, 0.01^2)$ and generated the rewards from (2.17), where we set $\mu = [-0.55, 0.666, -0.09, -0.232, 0.244, 0.55, -0.666, 0.09, 0.232, -0.244]^T$ and considered four cases for $\nu(t)$:

(i) $\nu(t) = 0$,

(ii) $\nu(t) = -b_{a^*(t)}(t)^T \mu$,

(iii) $\nu(t) = \log(t+1)$,

(iv) $\nu(t) = -\log(t+1)$.

We conducted 30 simulations in total for each case. The following graphs plot the cumulative regret $R(t)$ according to time t incurred by applying the 3 algorithms on the synthetic data. The solid lines represent the median values and the dashed lines represent the lower and upper 25% percentiles. Note that all 3 algorithms have a tuning parameter v that controls the degree of exploration. For each algorithm, we used the value of v which incurred minimum median regret over 30 simulations. These values were found by grid search.

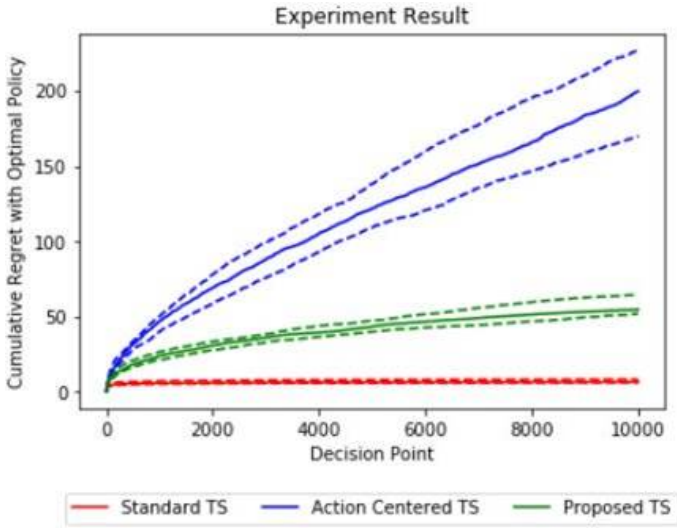

Figure 4.1: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations in case (i).

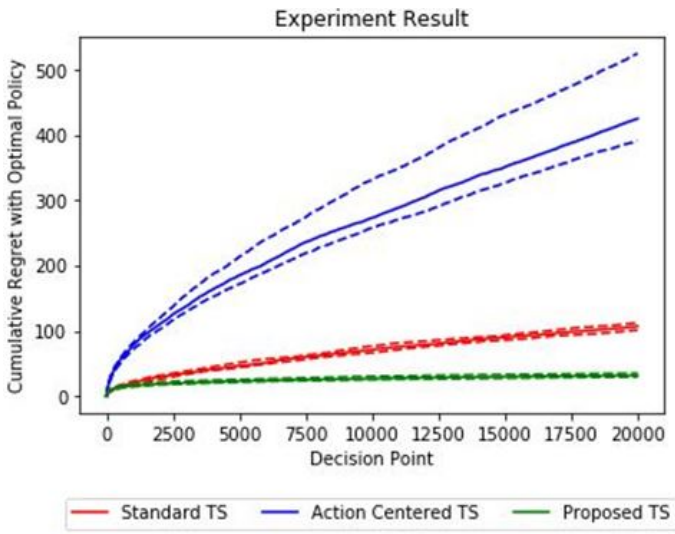

Figure 4.2: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations in case (ii).

In case (i), the original TS algorithm achieves lowest cumulative regret. However, in all 3 cases where $\nu(t)$ changes with time, the curve of the proposed method is the lowest. In these cases, the $R(t)$ of the original TS algorithm which assumes $\nu(t) = 0$ increases either constantly (case (ii) and (iv)) or exponentially (case (iii)) over time, indicating that the algorithm does not learn at all. On the other hand, although the method of Greenewald et al. (2017) has initially larger $R(t)$ than the TS algorithm, the slope of the graph slowly decreases over time in all four cases, showing that the algorithm indeed learns. However, due to the aforementioned constant $M = 1/\{p_{min}(1 - p_{max})\}$, the learning speed is significantly lower than ours.

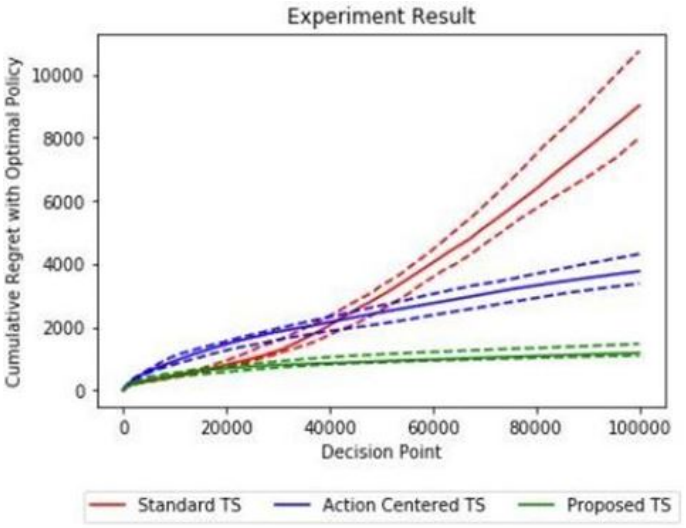

Figure 4.3: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations in case (iii).

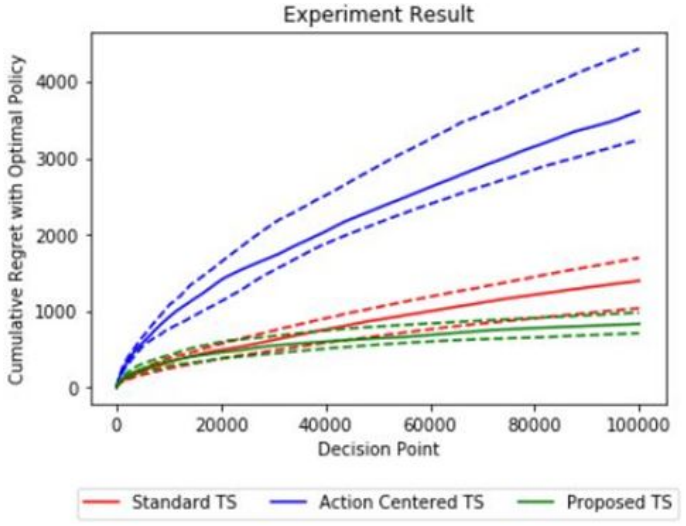

Figure 4.4: Median (solid), 1st and 3rd quartiles (dashed) of cumulative regret over 30 simulations in case (iv).

Chapter 5

Real data analysis

We applied the proposed method and existing methods to the R6A dataset provided by Yahoo! Webscope. The data is observational log data of user clicks from May 1st, 2009 to May 10th, 2009, which corresponds to 45,811,883 user visits. At every visit, one article was chosen uniformly at random from 20 articles ($N = 20$) and was displayed in the Featured tab of the Today module on Yahoo! front page (Figure 5.1). The reward $r_i(t)$ is binary, taking value 1 if the visiting user clicked the i-th article, and $r_i(t) = 0$ otherwise. For each article i, there is a context vector $b_i(t) \in \mathbb{R}^{35}$, which is constituted of 5 extracted user features, 5 extracted article features and their products. The extracted features were constructed from high-dimensional raw user and article features using a dimension reduction method of Chu et al. (2009).

As we have mentioned in Chapter 1, retrospective evaluation of a new reinforcement learning policy using observational real data calls for off-policy evaluation methods. This is because in observational data, the rewards of the actions that were not chosen by

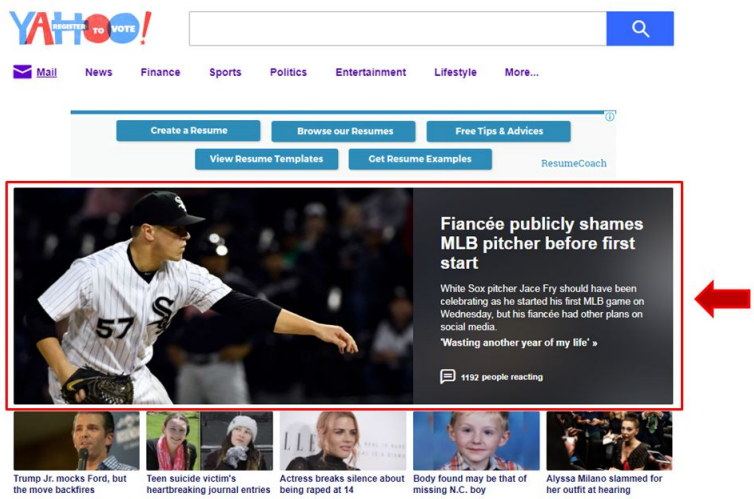

Figure 5.1: Yahoo! Featured tab screenshot image

the original policy are missing, so the evaluation of a new policy cannot be done straightforwardly since a new policy would make different action choices from the original policy. In the Yahoo! Webscope data, only the rewards of the displayed articles are observed. The displayed articles were chosen by the uniform random policy, which is totally different from the algorithms that we want to evaluate.

In Section 5.1, we review the off-policy evaluation method of Li et al. (2011) which enables unbiased estimation of the total reward of any policy when the action choice probabilities of the original logging policy are known. Then in Section 5.2, we apply the method of Li et al. (2011) to evaluate the proposed algorithm and other existing algorithms.

5.1 Off-policy evaluation method

Denote the policy that we want to evaluate as A. Based on observed trials ($\mathcal{H}_{t-1}$) and the current context $\boldsymbol{b}(t) = \{b_i(t)\}_{i=1}^N$, A chooses an arm $a(t) = \mathtt{A}(\mathcal{H}_{t-1}, \boldsymbol{b}(t))$ and receives reward $r_{a(t)}(t)$. The goal is to estimate the total T-trial reward of A,

$$G_{\mathtt{A}}(T) := \mathbb{E}\Big[\sum_{t=1}^{T} r_{a(t)}(t)\Big],$$

using data (S) collected from a different logging policy L.

5.1.1 Assumptions

The off-policy evaluation method of Li et al. (2011) requires some mild assumptions on the data S. We first present the assumptions.

- Collected data S is a *sufficiently long* stream of events $(\boldsymbol{b}, a, r_a)$, where $\{\boldsymbol{b}, \boldsymbol{r}\} \overset{i.i.d.}{\sim} D$ and a is chosen according to policy L.
- The logging policy L is a *randomized* policy, selecting each arm with positive probability.

We note that the first assumption does not cover the case where $\nu(t)$ is adaptive to the past trials. Still, the conditional distribution of $\nu(t)$ given $\boldsymbol{b}(t)$ is not restricted.

5.1.2 Algorithm : when L selects each arm uniformly at random.

The first algorithm presented by Li et al. (2011) provides a method to unbiasedly estimate $G_A(T)$ when the logging policy L is a uniform random policy. We present the algorithm and the theorem

which shows that $\mathbb{E}(\hat{G}_A(T)) = G_A(T)$ for any policy $\mathtt{A}$. As will be shown in Section 5.1.3, this method can be easily extended to the case where $\mathtt{L}$ is not a uniform random policy but the action choice probabilities are known and are positive.

Algorithm 6 Policy Evaluator (Li et al., 2011)

Inputs: $T > 0$; policy $\mathtt{A}$; stream of events S
$\mathcal{H}_0 \longleftarrow \emptyset$ {An initially empty history}
$\hat{G}_{\mathtt{A}} \longleftarrow 0$ {An initially zero total reward}
for $t = 1, 2, \cdots, T$ **do**
    repeat
        Get next event $\left(\boldsymbol{b}, a, r_a\right)$ from S
    until $\mathtt{A}(\mathcal{H}_{t-1}, \boldsymbol{b}) = a$
    $\mathcal{H}_t \longleftarrow \text{CONCATENATE}(\mathcal{H}_{t-1}, (\boldsymbol{b}, a, r_a))$
    $\hat{G}_{\mathtt{A}} \longleftarrow \hat{G}_{\mathtt{A}} + r_a$
end for
Output: $\hat{G}_{\mathtt{A}}$

Theorem 5.1.1. *Suppose the logging policy* $\mathtt{L}$ *selects each arm uniformly at random. Then for all distributions* D*, all algorithms* $\mathtt{A}$*, all* T*, all sequences of events* $\mathcal{H}_T$*, and all stream* S *satisfying the assumptions, we have*

$$\Pr_{\texttt{Policy Evaluator}(A,S)}\left(\mathcal{H}_T\right) = \Pr_{A,D}\left(\mathcal{H}_T\right),$$

i.e., the distribution of $\mathcal{H}_T$ *retained by the* `Policy Evaluator` *and the distribution of* $\mathcal{H}_T$ *obtained by applying policy* $\mathtt{A}$ *on real-world events from* D *are equivalent.*

Proof. Proof is done by induction on $t = 1, \cdots, T$. Suppose

$$\Pr_{\texttt{Policy Evaluator(A,}S\texttt{)}}\big(\mathcal{H}_{t-1}\big) = \Pr_{\texttt{A},D}\big(\mathcal{H}_{t-1}\big).$$

We want to prove the same statement for any history $\mathcal{H}_t$. We only need to show

$$\Pr_{\texttt{Policy Evaluator(A,}S\texttt{)}}\big((\boldsymbol{b}(t), a, r_a(t)) \mid \mathcal{H}_{t-1}\big) = \Pr_{\texttt{A},D}\big((\boldsymbol{b}(t), a, r_a(t)) \mid \mathcal{H}_{t-1}\big), \tag{5.1}$$

where $(\boldsymbol{b}(t), a, r_a(t))$ is the t-th event of $\mathcal{H}_t$. Note that

$$\begin{aligned}
(RHS \text{ of } (1)) &= \Pr_{\texttt{A},D}\big((\boldsymbol{b}(t), r_a(t))|\mathcal{H}_{t-1}\big)\Pr_{\texttt{A},D}\big(a \mid \boldsymbol{b}(t), r_a(t), \mathcal{H}_{t-1}\big) \\
&= \Pr_{D}\big((\boldsymbol{b}(t), r_a(t))\big)\Pr_{\texttt{A}}\big(a \mid \boldsymbol{b}(t), \mathcal{H}_{t-1}\big) \\
&= \Pr_{D}\big((\boldsymbol{b}(t), r_a(t))\big)\Pr\big(\texttt{A}(\mathcal{H}_{t-1}, \boldsymbol{b}(t)) = a\big),
\end{aligned}$$

and

$$\begin{aligned}
(LHS \text{ of } (1)) = &\Pr_{\texttt{Policy Evaluator(A,}S\texttt{)}}\big((\boldsymbol{b}(t), r_a(t))|\mathcal{H}_{t-1}\big)\times \\
&\Pr_{\texttt{Policy Evaluator(A,}S\texttt{)}}\big(a \mid \boldsymbol{b}(t), r_a(t), \mathcal{H}_{t-1}\big) \\
= &\Pr_{\texttt{Policy Evaluator(A,}S\texttt{)}}\big((\boldsymbol{b}(t), r_a(t))|\mathcal{H}_{t-1}\big)\times \\
&\Pr\big(\texttt{A}(\mathcal{H}_{t-1}, \boldsymbol{b}(t)) = a\big).
\end{aligned}$$

Meanwhile, in the `Policy Evaluator` algorithm, the probability of exiting the loop does not depend on $\mathcal{H}_{t-1}$ nor the policy `A` nor

the current context $\boldsymbol{b}$, because

$$\begin{aligned}\Pr_{\texttt{Policy Evaluator(A},S)}(\text{exit loop}|\mathcal{H}_{t-1}) &= \Pr(\texttt{A}(\mathcal{H}_{t-1},\boldsymbol{b})=a)\\ &= \sum_{i=1}^{N}\Pr(\texttt{A}(\mathcal{H}_{t-1},\boldsymbol{b})=i)\Pr_{\texttt{L}}(a=i)\\ &= \sum_{i=1}^{N}\Pr(\texttt{A}(\mathcal{H}_{t-1},\boldsymbol{b})=i)\frac{1}{N}\\ &= \frac{1}{N}. \qquad (5.2)\end{aligned}$$

Therefore, the probability that the t-th event of $\mathcal{H}_t$ is $(\boldsymbol{b}(t), r_a(t))$ does not depend on $\mathcal{H}_{t-1}$ nor $\texttt{A}$, i.e.,

$$\Pr_{\texttt{Policy Evaluator(A},S)}\big((\boldsymbol{b}(t), r_a(t))|\mathcal{H}_{t-1}\big) = \Pr_{D}\big((\boldsymbol{b}(t), r_a(t))\big). \quad (5.3)$$

□

5.1.3 Algorithm 2 : when L does not select each arm uniformly at random.

If the logging policy $\texttt{L}$ does not select each arm uniformly at random, the third equality of (5.2) will not hold in the above proof. Hence, the probability of exiting the loop will depend on $\mathcal{H}_{t-1}$, $\texttt{A}$, and $\boldsymbol{b}$, and the distribution of the t-th event of $\mathcal{H}_t$ will not follow D anymore. However, if $\texttt{L}$ selects each arm with positive probability, we can use rejection sampling to make (5.2) hold and thus (5.3) hold as well, at the cost of decreased data efficiency. Then the statement of Theorem 5.1.1 will hold, enabling unbiased estimation of $G_{\texttt{A}}$.

Theorem 5.1.2. *Suppose the logging policy* $\texttt{L}$ *selects each arm with positive probability. Specifically, suppose* $\exists M \geq 1$ *such that*

Algorithm 7 `Policy Evaluator2`

Inputs: $T > 0$; policy $\mathtt{A}$; stream of events S; constant $M \geq 1$ which satisfies $(1/N) \leq M\Pr(\mathtt{A}(\mathcal{H}_\tau, \boldsymbol{b}) = i)$ for every history $\mathcal{H}_\tau$, every $\boldsymbol{b}$ and every arm i.
$\mathcal{H}_0 \longleftarrow \emptyset$ {An initially empty history}
$\hat{G}_{\mathtt{A}} \longleftarrow 0$ {An initially zero total reward}
for $t = 1, 2, \cdots, T$ **do**
    repeat
        repeat
            Get next event $\big(\boldsymbol{b}, a, r_a\big)$ from S
        until $\mathtt{A}(\mathcal{H}_{t-1}, \boldsymbol{b}) = a$
        Generate $U \sim \text{Uniform}(0, 1)$.
    until $U \leq (1/N)/M/\Pr(\mathtt{A}(\mathcal{H}_{t-1}, \boldsymbol{b}) = a)$
    $\mathcal{H}_t \longleftarrow \text{CONCATENATE}(\mathcal{H}_{t-1}, (\boldsymbol{b}, a, r_a))$
    $\hat{G}_{\mathtt{A}} \longleftarrow \hat{G}_{\mathtt{A}} + r_a$
end for
Output: $\hat{G}_{\mathtt{A}}$

$(1/N) \leq M\Pr(\mathtt{A}(\mathcal{H}_\tau, \boldsymbol{b}) = i)$ *for every history* $\mathcal{H}_\tau$*, every* $\boldsymbol{b}$ *and every arm* i*. Then for all distributions* D*, all algorithms* $\mathtt{A}$*, all* T*, all sequences of events* $\mathcal{H}_T$*, and all stream* S *satisfying the assumptions, we have*

$$\Pr_{\texttt{Policy Evaluator2}(\mathtt{A},S)}\big(\mathcal{H}_T\big) = \Pr_{\mathtt{A},D}\big(\mathcal{H}_T\big),$$

i.e., the distribution of $\mathcal{H}_T$ *retained by the* `Policy Evaluator2` *and the distribution of* $\mathcal{H}_T$ *obtained by applying policy* $\mathtt{A}$ *on real-world events from* D *are equivalent.*

Proof. The proof follows the lines of the proof of Theorem 5.1.1.

We just need to prove that the probability of exiting the loop is independent of $\mathcal{H}_{t-1}$, $\mathtt{A}$, and current context $\boldsymbol{b}$:

$$\begin{aligned}\Pr_{\texttt{Policy Evaluator2}(\mathtt{A},S)}(\text{exit loop}|\mathcal{H}_{t-1}) &= \Pr(\mathtt{A}(\mathcal{H}_{t-1},\boldsymbol{b})=a)\times \\ &\quad \Pr(U \le \frac{1}{NM\Pr(\mathtt{A}(\mathcal{H}_{t-1},\boldsymbol{b})=a)}) \\ &= \frac{1}{NM}.\end{aligned}$$

Therefore, the probability that the t-th event of $\mathcal{H}_t$ is $(\boldsymbol{b}(t), r_a(t))$ is independent of $\mathcal{H}_{t-1}$ and $\mathtt{A}$, i.e.,

$$\Pr_{\texttt{Policy Evaluator2}(\mathtt{A},S)}\big((\boldsymbol{b}(t), r_a(t))|\mathcal{H}_{t-1}\big) = \Pr_{D}\big((\boldsymbol{b}(t), r_a(t))\big).$$

□

5.2 Application results

We evaluated the uniform random policy, TS algorithm and the proposed algorithm using the Yahoo! Webscope data. Since the original logging policy was a uniform random policy, we used Algorithm 6 to unbiasedly estimate the total rewards of each algorithm. We used data of May 1st, 2009 as tuning data to choose the optimal exploration parameter v for the TS algorithm and the proposed algorithm, respectively. Then we conducted main analysis on data from May 2nd, 2009 to May 10th, 2009.

Recall that Algorithm 6 can retain only $\frac{1}{N} = \frac{1}{20}$ of the whole data from May 2nd to May 10th, 2009. This corresponds to $T = 1900000$. We fixed the value of T to $T = 1900000$ a priori, and conducted Algorithm 6 for 10 times on the same data for each

algorithm. Since the evaluated algorithms are all randomized algorithms, each of the 10 runs pick up different actions, giving 10 different estimates for each algorithm. We report the mean, 1st quartile and 3rd quartile of the estimates for each algorithm in Table 5.1.

Policies	Mean	1st Q.	3rd Q.
Uniform policy	66696.7	66515.0	66832.75
Thompson sampling	86907.0	85992.75	88551.25
Proposed algorithm	90689.7	90177.25	91166.25

Table 5.1: Mean, 1st quartile (1st Q.) and 3rd quartile (3rd Q.) of user clicks achieved by each algorithm over 10 runs

We verify that the contextual bandit algorithms achieve much more higher user click rates than the uniform random policy. Among the contextual bandit algorithms, the proposed algorithm which assumes a nonstationary nonparametric intercept term in the reward distribution increased the average user click rate by 4.4% compared to the original TS algorithm.

Chapter 6

Concluding remarks

This thesis proposes a new contextual MAB algorithm for a semi-parametric additive reward model. In this model, the distribution of the baseline reward is allowed to change with time in an arbitrary manner. On the other hand, it is assumed that the amount of variation in the reward due to a specific action is linear with respect to the context information of the action, which is a reasonable assumption. This model is well suited to realistic problems such as news article recommendation, web page ad placement algorithms and mobile healthcare systems, because the baseline rewards are bound to change in an unexpected manner and sometimes can adapt to the past.

The proposed algorithm improves on the methods of Greenewald et al. (2017) and Krishnamurthy et al. (2018) which address the same reward model. Using concentration inequalities for vector-valued martingales, we proved that the high-probability regret upper bound of our method matches that of the Thompson sampling algorithm for linear reward models. We also applied the

proposed and existing methods on both synthetic data and real, news article recommendation data, where the results showed that the proposed method is superior.

Bibliography

Abbasi-Yadkori, Y., Pál, D. and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems,* 2312–2320.

Abramowitz, M. and Stegun, I.A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Washington, DC: National Bureau of Standards.

Agrawal, S. and Goyal, N. (2013). Thompson sampling for contextual bandits with linear payoffs. *Proceedings of the 30th International Conference on Machine Learning*, 127–135.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research,* **3**, 397–422.

Bercu, B. and Touati, A. (2008). Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability,* **18**(5), 1848–1869.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R.E. (2011). Contextual bandit algorithms with supervised

learning guarantees. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 19–26.

Chu, W., Park, S.T., Beaupre, T., Motgi, N., Phadke, A., Chakraborty, S. and Zachariah, J. (2009). A case study of behavior-driven conjoint analysis on Yahoo!: front page today module. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining,* 1097-1104.

Chu, W., Li, L., Reyzin, L. and Schapire, R.E. (2011). Contextual bandits with linear payoff functions. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics,* 208–214.

Dani, V., Hayes, T. P. and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *Conference on Learning Theory*, 355–366.

de la Peña, V. H., Klass, M. J. and Lai, T. L. (2004). Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. *Annals of Probability,* **32**(3A), 1902–1933.

de la Peña ,V. H., Klass, M. J. and Lai, T. L. (2009). Theory and applications of multivariate self-normalized processes. *Stochastic Processes and their Applications,* **119**(12), 4210–4227.

Even-Dar, E., Mannor, S. and Mansour, Y. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research,* **7**, 1079–1105.

Ferreira, K., Simchi-Levi, D. and Wang, H. (2018). Online network revenue management using Thompson sampling. *Operations Research,* **66**(6), 1586–1602.

Greenewald, K., Tewari, A., Murphy, S. and Klasnja, P. (2017). Action centered contextual bandits. *Advances in Neural Information Processing Systems*, 5977–5985.

Kawale, J., Bui, H.H., Kveton, B., Tran-Thanh, L. and Chawla, S. (2015). Efficient Thompson sampling for online matrix-factorization recommendation. *Advances in Neural Information Processing Systems*, 1297–1305.

Krishnamurthy, A., Wu, Z. S. and Syrgkanis, V. (2018). Semiparametric contextual bandits. *Proceedings of the 35th International Conference on Machine Learning.*

Lai, T.L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics,* **6**(1), 4–22.

Langford, J., Strehl, A. and Wortman, J. (2008). Exploration scavenging. *Proceedings of the 25th International Conference on Machine Learning,* 528–535.

Li, L., Chu, W., Langford, J. and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. *Proceedings of the 19th International Conference on World wide web*, 661–670.

Li, L., Chu, W., Langford, J. and Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommen-

dation algorithms. *Proceedings of the 4th ACM International Conference on Web search and data mining,* 297–306.

Neu, G. (2015). Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 3168–3176.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, **58**(5), 527–535.

Schwartz, E.M., Bradlow, E.T. and Fader, P.S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science,* **36**(4), 500–522.

Tewari, A., and Murphy, S.A. (2017). From ads to interventions: contextual bandits in mobile health. *Mobile Health*, 495–517. Springer, Cham.

Thompson, W.R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**(3/4), 285–294.

Wyatt, J. (1997). Exploration and inference in learning from reinforcement. Ph.D. thesis, Department of Artificial Intelligence, University of Edinburgh, Scotland.

Yahoo! Webscope. Yahoo! Front Page Today Module User Click Log Dataset, version 1.0. `http://webscope.sandbox.yahoo.com`. Accessed: 09/01/2019.

국문초록

다중 슬롯 머신 (MAB) 알고리즘은 순차 결정 문제를 다루는 연구 분야로서, 특정 환경 안에서 학습자에게 선택 가능한 다수의 행동들이 주어졌을 때, 이들 중 보상을 최대화하는 행동을 선택하는 방법론이다. 학습자는 행동을 선택하고 보상을 받는 과정을 반복하면서 보상 메커니즘에 대한 정보를 축적하고 학습하여 시간이 지남에 따라 최적의 행동에 가까운 행동을 선택하게 된다. 사이드 정보를 활용하는 다중 슬롯 머신 (Contextual MAB) 알고리즘은 순차적 의사 결정 시에 사이드 정보를 활용하는 MAB 알고리즘이며 최근 Yahoo!의 뉴스 기사 추천 시스템에 적용되어 기존에 비해 기사 클릭수를 크게 증가시키면서 많은 성과를 거두었다. 이 외에도 Contextual MAB 알고리즘이 주로 이용되는 분야로는 웹 페이지 광고 배치 알고리즘, 수익 관리, 모바일 헬스 시스템 등 다양하다. 더 좋은 MAB 알고리즘은 더 많은 보상과 수익을 창출할 수 있다는 점에서 매우 중요한 연구 분야다. 그러나 현재까지 제안된 대부분의 Contextual MAB 알고리즘은 보상과 사이드 정보 사이에 제한적인 선형 모형을 가정한다. 특히 보상 값의 분포가 시간에 따라 변하지 않는 다는 가정은 앞서 소개한 실제 문제들에 적용하기에는 비현실적이라는 지적을 받는다. 본 논문에서는 선형 가정보다 완화된 준모수적 가법 모형 하에서도 좋은 성능을

가지는 새로운 Contextual MAB 알고리즘을 제안한다. 준모수적 보상 모형 하에서 제안된 알고리즘에 의해 발생되는 누적 보상이 최적 보상으로 수렴하는 속도는 더 제한적인 선형 모형 하에서 톰슨 샘플링 알고리즘에 의해 발생되는 누적 보상이 수렴하는 속도와 유사하다. 또한, 제안된 방법은 동일한 모형을 다루는 두개의 선행 연구에 비해 덜 제한적이고 구현하기 쉬우며, 구현 속도가 더 빠르다. 시뮬레이션을 통해 제안된 알고리즘과 기존 알고리즘의 표본 성질을 비교한 결과를 소개한다. 더불어, Yahoo! 웹스코프가 제공하는 Yahoo! 뉴스 기사 추천 로그 데이터에 제안된 방법을 적용한 결과도 소개한다.

주요어 : 다중 슬롯 머신 알고리즘, 순차적 결정, 톰슨 샘플링, 준모수적 가법 모형

학 번 : 2015 – 30089